

# Model Selection for Likelihood-free Bayesian Methods Based on Moment Conditions: Theory and Numerical Examples

Cheng Li\* and Wenxin Jiang<sup>†</sup>

Department of Statistics, Northwestern University

## Abstract

An important practice in statistics is to use robust likelihood-free methods, such as the estimating equations, which only require assumptions on the moments instead of specifying the full probabilistic model. We propose a Bayesian flavored model selection approach for such likelihood-free methods, based on (quasi-)posterior probabilities from the Bayesian Generalized Method of Moments (BGMM). This novel concept allows us to incorporate two important advantages of a Bayesian approach: the expressiveness of posterior distributions and the convenient computational method of MCMC. Many different applications are possible, including modeling the correlated longitudinal data, the quantile regression, and the graphical models based on partial correlation. We demonstrate numerically how our method works in these applications. Under mild conditions, we show that theoretically the BGMM can achieve the posterior consistency for selecting the unknown true model, and that it possesses a Bayesian version of the oracle property, i.e. the posterior distribution for the parameter of interest is asymptotically normal and is as informative as if the true model were known. In addition, we show that the proposed quasi-posterior is valid to be interpreted as an approximate conditional distribution given a data summary.

---

\*Email: chengli2014@u.northwestern.edu.

<sup>†</sup>Email: wjiang@northwestern.edu.

*Key words:* Bayesian, GEE (generalized estimating equations), GMM (generalized method of moments), MCMC, model selection, moment condition, oracle property, posterior.

# 1 Introduction

## 1.1 Motivation

As the title of our paper suggests, there are three aspects of our work: a moment based method, a Bayesian flavored treatment, and model selection. In this subsection, we will first briefly explain why we are motivated to address all these three aspects in this paper, or in other words, what are the corresponding advantages that we see in this approach. In later subsections, we will review the related literature and formulate the method that we study.

1. *A moment based approach.* Our paper follows a long-standing statistical tradition, where people try to use moment based methods, such as GEE (generalized estimating equations in biostatistics) or GMM (generalized method of moments in econometrics) to avoid assuming a complete probability model and a likelihood function. The advantage is that the inferential results are robust and often depend only on the assumptions of the first and second order moments, instead of the complete probabilistic distribution.

2. *A Bayesian flavored approach.* Our moment-based method is Bayesian flavored. A quasi-posterior can be derived from a prior distribution and a quasi-likelihood constructed from the GMM criterion function, which has a simple explicit analytic form, compared to an alternative approach of a Bayesian nonparametric likelihood, and still has a valid interpretation as an approximate posterior given a data summary. This accommodates two important advantages of the Bayesian approach: the expressiveness of posterior distributions and the convenient computational method of MCMC (Markov chain Monte Carlo). These are particularly useful for the model selection problem that we study. We are able to report the most probable model, the second most probable model, etc., together with their quasi-posterior probabilities. We can also use the reversible jump MCMC algorithm (Green 1995, Dellaportas et al. 2002) to traverse the space of different models to simulate the quasi posterior probabilities.

3. *A model selection approach.* Our main goal is to study the theoretical properties and applications of model selection using our proposed quasi-Bayesian posterior. By doing model selection, we search for a simpler model that can describe the data well. This is useful for a parsimonious understanding of the data, and also useful for improving the predictive performance. Here we further illustrate this second point in the simple linear regression context with  $p$  regressors. Suppose the data can be described by a simple submodel with only  $k$  regressors ( $k < p$ ). The predictive performance of this submodel, such as the predictive mean square error (pMSE), can be defined as the oracle performance, since this true submodel is not known in practice. In general, the pMSE of this  $k$ -variable submodel scales as  $k/n$  for sample size  $n$ , while the pMSE under the  $p$ -variable full model scales as  $p/n$ , which increases linearly with the number of regressors  $p$ , even if most of the  $p$  regressors are irrelevant. Our theoretical results imply that our proposed model selection method asymptotically achieves the predictive performance from the oracle submodel, which is therefore much better than the method without model selection.

## 1.2 Method and related works

We consider the estimation problem based on the unconditional moment restrictions

$$E[g(D, \theta)] = 0 \tag{1}$$

where  $D$  is a set of random variables with domain  $\mathcal{D}$ ,  $\theta$  is a  $p$ -dimensional vector of parameters to be estimated, and  $g$  is a  $m$ -dimensional mapping from  $\mathcal{D} \times \mathbb{R}^p$  to  $\mathbb{R}^m$ . Typically it is necessary to have  $m \geq p$  for the point identification of  $\theta$ . Given an i.i.d. or stationary realization  $\mathbf{D} = \{D_1, D_2, \dots, D_n\}$  of  $D$ , one can either fully specify the underlying data generating process of  $D$ , or estimate  $\theta$  directly from such a set of  $m$  unconditional moments. This moment based estimation problem is important and has been extensively studied in econometrics and statistics. Well known methods include the generalized method of moments (GMM, Hansen 1982, Hansen et al. 1996, Newey 2004), the empirical likelihood (EL, Owen 1988, Qin and Lawless 1994), the exponential tilting (ET, Kitamura and Stutzer 1997), the exponential tilted empirical likelihood (ETEL, Schennach 2005, 2007) and the generalized empirical likelihood (GEL, Newey and Smith 2004). Essentially they all share the same first order efficiency of optimally weighted GMM estimator, and have been applied to independent data, time series data and panel data in econometrics. On the other hand, researchers in statistics also use the moment

based methods for constructing efficient estimators, especially for clustered and correlated longitudinal data. For example, Qu et al. (2000) proposed a GMM type estimator to avoid the inefficiency from misspecified working correlation matrices in generalized estimating equations (GEE) for longitudinal data. Wang et al. (2010) considered the EL approach to address the within-subject correlation structure. In general, the moment based estimation methods only require information on the low order moments of  $D$  and are therefore more flexible, efficient and robust to model misspecification, as long as the moment conditions are correctly specified.

In this paper, we consider the case where in (1), the dimension  $p$  of parameter  $\theta$  can possibly diverge with the sample size  $n$ , while the true parameter  $\theta_0$  could possibly lie in a lower dimensional subspace. Hereafter without causing confusion, we suppress the dependence of  $p$  on  $n$  and work with the triangular sequence of models. The increasing dimensionality of the parameter space arises from a variety of statistical applications. For example, in clinical trial studies, many variables related to a certain disease are measured and recorded longitudinally, so the number of covariates and their interaction effects can be reasonably large compared to the sample size. In financial economics, certain stock index could be possibly affected by a large volume of related stock prices, underlying economic variables and their lagged values. In genetic network models, the number of edges in the network is combinatorially large given a set of variables such as DNA copy numbers, RNA copy numbers and protein expressions. Therefore for the model interpretation and the accuracy of prediction, it is reasonable to assume that the number of truly relevant variables is relatively smaller subset of the original pool. Our goal is to consistently select these relevant variables and estimate their effects, namely the nonzero components of  $\theta_0$ , when the specification of full probabilistic model is unavailable but a sufficient number of moment conditions are present.

Moment based estimation methods with a diverging number of parameters have been studied extensively in frequentists' literature, pioneered by the early seminal work for the asymptotic behavior of M-estimation in Huber (1973) and Portnoy (1984, 1985). Recently Wang (2011) considered the estimating equations for correlated longitudinal data, with a focus on the binary response, and showed that without model selection, the estimated parameter is consistent if  $p^2/n \rightarrow 0$  and asymptotically normal if  $p^3/n \rightarrow 0$ . For the model selection problem, Wang et al. (2012) incorporated the SCAD penalty into the

estimating equations, and showed that the solution can consistently select the nonzero parameters  $\theta$  and the nonzero components are asymptotically normal. Similar oracle results are established in Li et al. (2011) with SCAD penalty for general M-estimators, and in Li et al. (2013) for the smooth-threshold GEE with nuisance parameters taken into account. For possibly over-identified unconditional moments with  $m > p$  and  $m, p$  in the same order, Caner and Zhang (2013) explored the GMM estimator with the elastic net penalty and obtained the model selection consistency and asymptotic normality under  $p^3/n \rightarrow 0$ . Leng and Tang (2012) considered the EL method and showed that for a class of penalty functions, the EL estimator shares similar oracle properties under  $p^5/n \rightarrow 0$ . Cho and Qu (2013) imposed the SCAD penalty on the quadratic inference function instead of on the empirical estimating equations in Wang et al. (2012) and obtained oracle properties under  $p^4/n \rightarrow 0$ .

Our work focuses on the Bayesian inference of  $\theta$  under moment constraint (1) with  $p$  diverging with  $n$ . Compared to the various frequentist regularization methods, the Bayesian literature on this problem still remains limited. One difficulty that hinders the fully probabilistic Bayesian modeling is that some prior distribution on both the distribution of  $D$  (denoted as  $P_D$ ) and the parameter  $\theta$  need to be specified, such that the pair  $(P_D, \theta)$  satisfies the set of restrictions (1). Recent progress in this direction includes Kitamura and Otsu (2011) and Florens and Simoni (2012). Kitamura and Otsu (2011) tried to minimize the Kullback-Leibler divergence of  $P_D$  to a Dirichlet process, which leads to an ET type likelihood function that computationally requires optimizations within each MCMC iteration step. Florens and Simoni (2012) exploited the Gaussian process prior and required a functional transformation of the data that is only asymptotically Gaussian, which still gives a misspecified likelihood function in finite samples. Besides, both methods have been tested only on simple examples that involve a few parameters and moments, and their behavior under increasing dimensionality has not been fully studied. Instead, another possible simpler Bayesian way for modeling (1) is to construct the simple quasi-likelihood function

$$q(\mathbf{D}|\theta) = \frac{1}{\det\{2\pi\mathbf{V}_n/n\}^{\frac{1}{2}}} \exp \left\{ -\frac{n}{2} \bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta) \right\}, \quad (2)$$

where  $\bar{g}(\mathbf{D}, \theta)$  is the sample average of  $g(D_i, \theta)$ ,  $i = 1, 2, \dots, n$ ,  $\mathbf{V}_n$  is a  $m \times m$  positive definite matrix that could possibly depend on the data  $\mathbf{D}$ , and  $\det(\mathbf{A})$  denotes the determinant of a matrix  $\mathbf{A}$ . Here and after we use the symbol “ $q$ ” to denote the quasi-likelihood

function and the quasi-posterior. This quasi-likelihood function has been studied under a Bayesian framework in Kim (2002) and is named *the limited information likelihood (LIL)*, which minimizes the Kullback-Leibler divergence of the true data generating process  $P_D$  to the set of all distributions satisfying the less restrictive asymptotic constraint  $\lim_{n \rightarrow \infty} E[n\bar{g}(\mathbf{D}, \theta_0)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta_0)]/m = 1$ . This relation holds when we choose  $\mathbf{V}_n$  to be a consistent estimator of the covariance matrix  $\text{Var}(g(\mathbf{D}, \theta_0))$ . Given a prior distribution  $\pi(\theta)$ , the quasi-posterior takes the form

$$q(\theta|\mathbf{D}) \propto \frac{1}{\det\{2\pi\mathbf{V}_n/n\}^{\frac{1}{2}}} \exp\left\{-\frac{n}{2}\bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta)\right\} \pi(\theta). \quad (3)$$

By using  $q(\mathbf{D}|\theta)$  in a Bayesian model, we only need to specify a prior on  $\theta$  and thus circumvent the difficulty of directly assigning a prior on the pair  $(P_D, \theta)$  with constraints (1). In the computational aspect,  $q(\mathbf{D}|\theta)$  takes an explicit analytical form that allows straightforward MCMC updating for the corresponding Bayesian posterior without any iterative optimization steps. Furthermore, when  $\mathbf{V}_n$  is chosen as a consistent estimator of  $\text{Var}(g(\mathbf{D}, \theta_0))$ , the exponential part of  $q(\mathbf{D}|\theta)$  resembles the optimally weighted GMM criterion function (Hansen 1982), which in large sample, can be viewed as a second order approximation to the true negative log-likelihood function that follows a chi-square distribution with  $p$  degrees of freedom if  $m = p$  and both are fixed (Yin 2009). We use the same terminology as Yin (2009) and call this *the Bayesian generalized method of moments* (BGMM). Below are some motivating examples that involve the unconditional moment conditions (1) and can be easily incorporated into the BGMM framework.

### 1.3 Three examples

**Example 1. Correlated longitudinal data.** In longitudinal studies, suppose the  $j$ th observation for the  $i$ th subject is the scalar response variable  $Y_{ij}$  and  $p$ -dimensional covariate vector  $X_{ij}$ . For simplicity, we assume that each subject has the same number of observations, i.e.  $j = 1, 2, \dots, s$  and  $i = 1, 2, \dots, n$ . Let  $Y_i = (Y_{i1}, \dots, Y_{is})^\top$ ,  $\mathbf{X}_i = (X_{i1}, \dots, X_{is})^\top$ , and  $E(Y_i|\mathbf{X}_i) = \mu_i(\theta)$ , where  $\mu_i(\theta) = (\mu(X_{i1}^\top\theta), \dots, \mu(X_{is}^\top\theta))^\top$  and  $\mu(\cdot)$  is a monotone link function. To account for the heteroscedasticity, we assume the conditional variance of  $Y_{ij}$  given  $X_{ij}$  is a function of the single index  $X_{ij}^\top\theta$ , i.e.  $\text{Var}(Y_{ij}|\mathbf{X}_i) = \phi(X_{ij}^\top\theta)$ . Then the frequentist GEE method estimates  $\theta$  by solving equations

$$n^{-1} \sum_{i=1}^n \frac{\partial \mu_i(\theta)^\top}{\partial \theta} \mathbf{S}_i^{-1} (Y_i - \mu_i(\theta)) = 0 \quad (4)$$

where  $\mathbf{S}_i = \mathbf{A}_i^{1/2} \mathbf{R} \mathbf{A}_i^{1/2}$ ,  $\mathbf{A}_i = \mathbf{A}_i(\theta) = \text{diag} \{ \phi(X_{i1}^\top \theta), \dots, \phi(X_{is}^\top \theta) \}$  is the diagonal matrix with the conditional variance of  $Y$  given  $X$  and  $\mathbf{R}$  is a working correlation matrix. Then if we denote the data as  $D_i = (Y_i, \mathbf{X}_i)^\top$ , then the moment is defined by

$$g(D_i, \theta) = \frac{\partial \mu_i(\theta)^\top}{\partial \theta} \mathbf{S}_i^{-1} (Y_i - \mu_i(\theta)). \quad (5)$$

Then the moment condition (1) is satisfied. The frequentist asymptotic normality and model selection consistency under diverging  $p$  and fixed  $s$  for this estimating equation has been studied in Wang (2011) and Wang et al. (2012). We will apply BGMM to the same setup and derive similar oracle properties.

**Example 2. Quantile regression.** Suppose that  $Y$  is a continuously distributed response variable, and  $X$  is a  $p$ -dimensional predictor vector for the  $\tau$ -th quantile ( $\tau \in (0, 1)$ ) of  $Y$ . The conditional quantile function of  $Y$  given  $X$  is specified by  $F_{Y|X}^{-1}(\tau) = X^\top \theta$ , where  $F_{Y|X}^{-1}$  is the inverse of conditional distribution function of  $Y$  given  $X$ . Then let  $D = (Y, X^\top)^\top$  and we can construct  $p$  unconditional moments as

$$g(D, \theta) = X [1(Y - X^\top \theta \leq 0) - \tau], \quad (6)$$

where  $1(\cdot)$  is the indicator function. The Bayesian method for quantile regression has been studied by, for example, Yu and Moyeed (2001), Kottas and Gelfand (2001), Li et al. (2010), Lancaster and Jun (2010), Yang and He (2012), and many others. It is worth noting that in some of these literature, the working likelihood is constructed only based on the check function of the linear quantile regression, which is usually not a valid likelihood even in the approximation sense. We will use the asymptotically interpretable LIL and derive the model selection results for quantile regression under increasing  $p$ , as a direct extension of Chernozhukov and Hong (2003), and also as a first order asymptotic equivalent to the Bayesian EL and ETEL methods from Yang and He (2012) and Lancaster and Jun (2010).

**Example 3. Partial correlation selection.** The partial correlation structure of a  $s$ -dimensional random vector  $Y$  is specified by its precision matrix  $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ , where  $\mathbf{\Sigma} = E[(Y - EY)(Y - EY)^\top]$  is the covariance matrix of  $Y$ . Hereafter without loss of generality, we assume that  $Y$  is centered such that  $EY = 0$ . The partial correlation between the  $i$ th and the  $j$ th components of  $Y$  is defined by  $\rho_{ij} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$ , where  $\omega_{ij}$  denotes the  $(i, j)$ th entry of  $\mathbf{\Omega}$ .  $\omega_{ij} = 0$  implies zero partial correlation between the  $i$ th and the  $j$ th components of  $Y$  given all the other components. For multivariate Gaussian random vector, there is an equivalence between the conditional independence and

the zero partial correlation. However, in the general case where multivariate Gaussian assumption is not satisfied, we can still use the second moment of  $Y$  to identify the zero entries in  $\Omega$ . Let  $\theta$  be the vectorized upper triangle part of  $\Omega$ . Then we can define the moment condition

$$g_{ij}(Y, \theta) = Y_i Y_j - (\Omega^{-1})_{ij}, \quad (7)$$

for  $1 \leq i \leq j \leq s$ , and the stacked moment vector  $g(Y, \theta)$  satisfies (1). We have  $\dim(\theta) = \dim(g) = \frac{s(s+1)}{2} =: p$  where  $\theta$  is just identifiable. The model selection problem for partial correlation has been studied in, for example, Drton and Perlman (2004), Jiang and Turnbull (2004), etc. We will consider the case where  $s$  (and hence  $p$  and  $\mathcal{M}$ ) increases with the sample size  $n$  and use BGMM to select the nonzeros out of the  $\frac{s(s-1)}{2}$  off-diagonal entries in  $\Omega$ .

## 1.4 Previous work on BGMM

The theoretical properties of BGMM has been investigated in Chernozhukov and Hong (2003), where they have shown that under fixed dimension  $p$ , the posterior mean of  $\theta$  has the same  $\sqrt{n}$  parametric rate as the usual GMM estimator, and the posterior distribution converges in total variation norm to normal. The latter result is known as the Bayesian central limit theorem (Bayesian CLT or Bernstein von Mises theorem), which has been generalized to BGMM increasing dimension  $p^4/n \rightarrow 0$  in Belloni and Chernozhukov (2009) for a wide class of moment conditions under an improper flat prior on  $\theta$ . Other applications of BGMM include the moment inequality models (Liao and Jiang 2010) and the nonparametric instrumental regression (Liao and Jiang 2011, Kato 2013). However, it is not trivial whether BGMM remains a feasible method for model selection with increasing  $p$ . We will answer this question and derive comparable results to frequentist literature such as Wang et al. (2012), Leng and Tang (2012), Cho and Qu (2013), etc.

## 1.5 Contributions of current paper

We extend Belloni and Chernozhukov (2009) further to the model selection problem with increasing  $p$  and derive oracle properties for the BGMM procedure. The detailed contributions of the current paper include the following:



1. We prove that BGMM automatically achieves the global model selection consistency under some regularity conditions on the moment function  $g(D, \theta)$  and the prior, with increasing rate  $p^4/n \rightarrow 0$  up to some logarithm factors. This is to say that the BGMM posterior probability of the true model converges to 1 with high probability.

2. We derive an oracle property for the BGMM procedure, which states that the BGMM posterior distribution converges in total variation norm to a normal distribution concentrated on the true model space with an efficient variance, as if the true model were known. This oracle property is the Bayesian analog of the frequentist post-model-selection oracle property of Fan and Li (2001). As far as we know, this is the first time that such an oracle property is formulated in a Bayesian context. We apply BGMM to our motivating examples and show that the model selection consistency and oracle property hold under mild regularity conditions on the data and the moments. To examine the empirical performance of BGMM, we implement the reversible jump MCMC algorithm for both simulation and real data examples, and demonstrate BGMM as a practically feasible and efficient alternative to the frequentist regularization methods.

3. We show that the BGMM quasi-posterior has a valid interpretation, as an approximate posterior conditional on a data summary that is equivalent to the GMM estimator. Particularly for the model selection, we derive the convergence rates of the Bayes factors for the BGMM method and the Bayesian method given the GMM estimator, and show that they have similar asymptotic behavior. Therefore, the model posterior probabilities from BGMM can be used directly for comparing different models.

## 1.6 Organization of the paper

The rest of the paper is organized as follows. In Section 2.2, we derive the oracle properties for BGMM model selection under increasing  $p$  based on a set of high level assumptions. We check these assumptions for the three previous motivating examples in Section 2.3. In Section 2.4, we discuss the validity of the proposed BGMM quasi-posterior. Section 3 provides numerical experiments and one real data analysis to illustrate the empirical performance of BGMM model selection. Section 4 includes further discussions.

## 1.7 Some useful notation

We define some useful notation. Let  $\|\cdot\|_k$  denote the  $L_k$  norm for  $k \in [0, \infty]$  and  $\|\cdot\|$  be the Euclidean norm ( $L_2$  norm). For any generic square matrix  $\mathbf{C}$ , let  $\underline{\lambda}(\mathbf{C})$ ,  $\bar{\lambda}(\mathbf{C})$  denote the smallest and the largest eigenvalues of a square matrix  $\mathbf{C}$ . Let  $\text{tr}(\mathbf{C})$  be the trace of  $\mathbf{C}$ . Let  $\|\mathbf{C}\| = \sqrt{\bar{\lambda}(\mathbf{C}^\top \mathbf{C})}$  be the matrix operator norm, and  $\|\mathbf{C}\|_F = \sqrt{\text{tr}(\mathbf{C}^\top \mathbf{C})}$  be the Frobenius norm of  $\mathbf{C}$ . For two stochastic sequence  $\{a_n\}$  and  $\{b_n\}$ , let  $a_n \prec b_n$ ,  $a_n \succ b_n$  and  $a_n \asymp b_n$  denote  $a_n = o(b_n)$ ,  $b_n = o(a_n)$  and  $a_n, b_n$  having the same order as  $n \rightarrow \infty$ .  $a_n \vee b_n = \max(a_n, b_n)$ . The notations  $o_p$  and  $O_p$  always refer to the probability measure  $P_{\mathbf{D}}$  of the sample  $\mathbf{D}$ . We use “ $C$ ” to denote any generic constant whose value can change in different places. “w.p.a.1” is an abbreviation for “with  $P_{\mathbf{D}}$  probability approaching 1”.

## 2 Some Theoretical Properties on Bayesian GMM Model Selection

The Bayesian model selection (BMS) problem has been extensively studied for Gaussian linear regression models and other generalized linear models. See for example, George and McCulloch (1993), Smith and Kohn (1996), Chipman et al. (2001), Ishwaran and Rao (2005), Liang et al. (2008) with a fixed number of predictors  $p$ , Jiang (2007), Moreno et al. (2010), Johnson and Rossell (2012), Liang et al. (2013) with high dimension  $p$ . In this paper, instead of imposing a specific probabilistic model assumption, we assume that the true parameter  $\theta_0$  can be point identified by some moment conditions (1) and possibly lie in a lower dimensional subspace of the whole parameter space  $\Theta \subseteq \mathbb{R}^p$ . When  $p$  is very large,  $\theta_0$  could be possibly sparse. We will also restrict  $\Theta$  to be a compact and connected set in  $\mathbb{R}^p$ , with finite  $L_2$  radius  $R = \sup_{\theta \in \Theta} \|\theta\|$  for some large constant  $R > 0$ .

Without loss of generality, in the following we will consider models generated by all the possible coordinate subspaces of  $\mathbb{R}^p$ , which leads to a total of  $2^p$  different models  $\mathcal{M}$  and the parameter space partition  $\Theta = \bigcup_{|\mathcal{M}| \leq p} \Theta(\mathcal{M})$ . Let  $k = |\mathcal{M}|$  ( $0 \leq k \leq p$ ) be the size of a generic model  $\mathcal{M}$ , which is the number of nonzero components in any  $\theta \in \mathcal{M}$ . Suppose  $\mathcal{M}_0$  is the true model space that contains  $\theta_0$ , and  $k_0 = |\mathcal{M}_0|$  is the dimension of  $\theta_0$ . For any generic  $\theta$ , let  $\theta = (\theta_1^\top, \theta_2^\top)^\top$  where  $\theta_1 \in \mathbb{R}^k$  and  $\theta_2 \in \mathbb{R}^{p-k}$  correspond to the nonzero and zero components respectively. So  $\theta_2 = 0$  if  $\theta \in \Theta(\mathcal{M})$ . We also emphasize

that the subscript “1” and “2” change with the model index  $\mathcal{M}$ , and in the following context,  $\theta_1$  is used as a dummy variable, referring to the nonzero components of any generic  $\theta$ .

For such a model selection setup, the prior distribution can be written in the hierarchical structure  $\pi(\theta) = \sum_{\mathcal{M}} \pi(\theta|\mathcal{M})\pi(\mathcal{M}) = \sum_{\mathcal{M},k} \pi(\theta|\mathcal{M})\pi(\mathcal{M}||\mathcal{M}| = k)\pi(k)$  for  $k = 0, 1, 2, \dots, p$ . If a model  $\mathcal{M}$  does not contain all the nonzero components for a given  $\theta$ , then  $\pi(\theta|\mathcal{M}) = 0$ . We assume that the parameter  $\theta$  can continuously take values and that each  $\pi(\theta|\mathcal{M})$  has a density function. For two different models  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , the (quasi-) Bayes factor of  $\mathcal{M}_1$  with respect to  $\mathcal{M}_2$  is defined as

$$\text{BF}_q[\mathcal{M}_1 : \mathcal{M}_2] = \frac{q(\mathbf{D}|\mathcal{M}_1)}{q(\mathbf{D}|\mathcal{M}_2)} = \frac{\int_{\Theta(\mathcal{M}_1)} q(\mathbf{D}|\theta, \mathcal{M}_1)\pi(\theta|\mathcal{M}_1)d\theta}{\int_{\Theta(\mathcal{M}_2)} q(\mathbf{D}|\theta, \mathcal{M}_2)\pi(\theta|\mathcal{M}_2)d\theta} \quad (8)$$

and accordingly the (quasi-) posterior odds is the product of the Bayes factor and the prior odds

$$\text{PO}_q[\mathcal{M}_1 : \mathcal{M}_2] = \frac{q(\mathbf{D}|\mathcal{M}_1)}{q(\mathbf{D}|\mathcal{M}_2)} \cdot \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} = \frac{\int_{\Theta(\mathcal{M}_1)} q(\mathbf{D}|\theta, \mathcal{M}_1)\pi(\theta|\mathcal{M}_1)d\theta}{\int_{\Theta(\mathcal{M}_2)} q(\mathbf{D}|\theta, \mathcal{M}_2)\pi(\theta|\mathcal{M}_2)d\theta} \cdot \frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)} \quad (9)$$

The BMS consistency we are going to establish is the global model selection consistency (Johnson and Rossell 2012), in the sense that asymptotically the true model  $\mathcal{M}_0$  will not only be the MAP model (*maximum a posteriori*) but also have posterior probability tending to 1. Equivalently, we will show that the sum of all posterior odds  $\text{PO}_q[\mathcal{M} : \mathcal{M}_0]$  with  $\mathcal{M} \neq \mathcal{M}_0$  converges to zero in probability. This strongest mode of consistency implies that the posterior mass will be concentrated around the true model and most of the  $2^p$  models receive negligible probabilities. This is a desirable property in practice for interpretation, since commonly used Bayesian estimation procedures such as model averaging will then involve only a few models instead of exponentially growing number of models under an increasing dimension  $p$ .

## 2.1 Assumptions

The set of assumptions below follows closely the set of conditions for Z-estimation in Belloni and Chernozhukov (2009). They are high level assumptions imposed on the data generation process, the model parameters, the moment conditions and the priors. They will be verified in Section 3 for all of our examples. For a specific model, these assumptions are not necessarily in the most general form, but they do cover a wide class of moment restricted models in practice and are sufficient for illustrating the theoretical

properties of BGMM.

For the data generation process, the true parameter  $\theta_0$ , and the growth rate of  $p$  and  $m$ , we make the following assumptions.

- *Assumption 1* (Data Generation Process)  $\{D_i, i = 1, 2, \dots, n\}$  is an i.i.d. sequence.  $Eg(D, \theta_0) = 0$  for some  $\theta_0 \in \Theta$ .  $\Theta$  is a compact and connected set with  $L_2$  radius  $R$  for some large constant  $R > 0$ , and it contains an open neighborhood of  $\theta_0$ .
- *Assumption 2* (Growth Rate) Let  $\dim(\theta) = p$ . Assume that  $p \leq m$  and  $p \asymp m$ ,  $p^4 \log^2 n/n \rightarrow 0$ ,  $p^{2+\alpha} \log n/n^\alpha \rightarrow 0$ , where  $\alpha$  is defined in Assumption 4.
- *Assumption 3* (Beta-min) Let  $\epsilon_n = \sqrt{p/n}$ . Assume  $1 \succeq \min_{j \in \mathcal{M}_0} |\theta_{0,j}| \succ \sqrt{\log n} \epsilon_n$ .

The i.i.d. assumption can be possibly relaxed to a weakly dependent stationary process using more involved techniques. The compactness assumption for the parameter space  $\Theta$  is standard, though in the increasing  $p$  setting, this could be possibly relaxed to a  $\Theta$  with a slowly diverging diameter  $R$ , with additional minor adjustments on the growth rate of  $p$ . The growth rate condition on  $p$  is the same as that in Belloni and Chernozhukov (2009), and should not be considered as restrictive because we are considering a very general class of models based on moment conditions, with no specific distributional assumptions the data (see also similar assumptions in Leng and Tang 2012). In fact, from our proofs, one can see that Assumption 2 is necessary to control the residual terms after a local linearization of the moment function  $g(D, \theta)$ , after applying the empirical process results from Belloni and Chernozhukov (2009) and Belloni et al. (2011). The beta-min condition here is technical and commonly used in frequentist high dimensional literature (see e.g. Bühlmann and van de Geer 2011, Wang et al. 2012, Leng and Tang 2012, Cho and Qu 2013, etc.) It gives the minimal magnitude of nonzero coefficients that could be detected by BGMM. When  $p$  has the order in Assumption 2, our minimal bound  $\sqrt{\log n} \epsilon_n$  decreases to zero fast, which is also less restrictive than the constant lower bound used in Johnson and Rossell (2012) for BMS.

The following assumptions are on the moment conditions.

- *Assumption 4* (Moment) (i) The moment function  $g(D, \theta)$  satisfies the continuity property  $\sup_{\eta \in \mathbb{R}^m, \|\eta\|=1} (E[(\eta^\top (g(D, \theta) - g(D, \theta_0)))^2])^{1/2} \leq O((\sqrt{p} \|\theta - \theta_0\|)^\alpha)$ , uniformly in  $\theta \in \Theta$  for some constant  $\alpha \in (0, 1]$ .
- (ii) The class of functions  $\mathcal{F} = \{\eta^\top (g(D, \theta) - g(D, \theta_0)), \theta \in \Theta, \eta \in \mathbb{R}^m, \|\eta\| = 1\}$

has an envelope function  $F$  almost surely bounded in  $L_2$  norm  $\|\cdot\|_{P_{D,2}}$  as order  $O(\sqrt{p})$ . The  $L_2$  uniform covering number  $N(\epsilon\|F\|_{P_{D,2}}, \mathcal{F}, L_2(P_D))$  satisfies

$$\log N(\epsilon\|F\|_{P_{D,2}}, \mathcal{F}, L_2(P_D)) = O\left(p \log\left(\frac{n}{\epsilon}\right)\right).$$

- *Assumption 5* (Linearization) (i)  $\|Eg(D, \theta)\| \geq \delta_0 \wedge \delta_1 \|\theta - \theta_0\|$  uniformly on  $\Theta$  for some positive constants  $\delta_0, \delta_1$ .
- (ii)  $\mathbf{G} := \nabla_{\theta} Eg(D, \theta_0)$  exists, and the eigenvalues of  $\mathbf{G}^{\top} \mathbf{G}$  are bounded from below and above as  $n \rightarrow \infty$ .
- (iii)  $\mathbf{H}(\theta) := \nabla_{\theta\theta^{\top}}^2 Eg(D, \theta)$  exists for  $\theta \in B_0(C\epsilon_n)$ , and uniformly over  $\theta \in B_0(C\epsilon_n)$  for any fixed  $C > 0$ ,  $\sup_{\|u\|=1, \|v\|=1, u, v \in \mathbb{R}^p} \|\mathbf{H}(\theta)[u, v]\| = O(\sqrt{p})$ .

Assumption 4 and 5 on moment function  $g(D, \theta)$  parallel the conditions ZE.1 and ZE.2 in Belloni and Chernozhukov (2009) respectively. The continuity index  $\alpha$  in Assumption 4(i) satisfies  $\alpha = 1$  for the mean regression, such as the examples of correlated longitudinal data and partial correlation selection, and  $\alpha = 1/2$  for the quantile regression model. The entropy condition in Assumption 4(ii) controls the complexity of the class of moment functions  $g(D, \theta)$ . Assumption 5(i) guarantees the point identification of the true parameter  $\theta_0$ , and part (ii) and (iii) impose mild assumptions on the first and second derivatives of  $Eg(D, \theta)$  around  $\theta_0$ . These regularity conditions are mainly used to derive large deviation bounds via empirical process results, and they will be verified later for our motivating examples. Note that unlike Wang et al. (2012), Leng and Tang (2012) and Cho and Qu (2013), we have not required the moment function  $g(D, \theta)$  itself to be differentiable. This allows more general applications to discontinuous  $g(D, \theta)$ , such as in the case of quantile regression.

- *Assumption 6* (Variance)  $\mathbf{V}_n$  is a positive definite matrix for all  $n$ , and converges in the matrix operator norm to  $\mathbf{V} = \text{Var}[g(D, \theta_0)]$ . The eigenvalues of  $\mathbf{V}_n$  and  $\mathbf{V}$  are bounded below and above for some positive constants  $\underline{\lambda}$  and  $\bar{\lambda}$  w.p.a.1.

Assumption 6 assumes that the positive definite weighting matrix  $\mathbf{V}_n$  is a consistent estimator of the covariance matrix of  $g(D, \theta)$  at  $\theta_0$ , similar to the preliminary estimator of the optimal weighting matrix used in the two step GMM estimation. Although this consistency of  $\mathbf{V}_n$  to  $\mathbf{V}$  is not required for the model selection consistency, it is necessary for the valid posterior inference such as posterior credible sets. Essentially  $\mathbf{V}_n$  needs to satisfy the generalized information inequality (Chernozhukov and Hong 2003), such that

the LIL asymptotically satisfies the second Bartlett identity as a true likelihood function does. Note that when  $p$  is of the order in Assumption 2, such consistent estimator  $\mathbf{V}_n$  usually exists for all our motivating examples.

Finally we impose the following assumptions on the prior.

- *Assumption 7* (Prior on  $\theta$ ) (i)  $\pi(\theta|\mathcal{M})$  has a density function restricted to  $\Theta$ , and is bounded above by a constant  $c_\pi$  uniformly over all model spaces  $\mathcal{M}$ .  
(ii)  $|\pi(\theta|\mathcal{M}) - \pi(\theta_{0,1}|\mathcal{M})| = o(1)$  holds uniformly over  $\theta \in \Theta(\mathcal{M})$  and  $\|\theta - \theta_{0,1}\| \leq C\epsilon_n$ , and uniformly for all models  $\mathcal{M}$ , all fixed  $C > 0$  and all sufficient large  $n$ , where  $\theta_{0,1}$  is the subvector of  $\theta_0$  restricted to  $\Theta(\mathcal{M})$ .  
(iii)  $\pi(\theta_0|\mathcal{M}) \geq e^{-c_0|\mathcal{M}|}$  for some constant  $c_0 > 0$  uniformly for all  $\mathcal{M} \supseteq \mathcal{M}_0$ .  
 $|\log \pi(\theta|\mathcal{M}_0) - \log \pi(\theta_0|\mathcal{M}_0)| = o(1)$  holds uniformly over  $\theta \in \Theta(\mathcal{M}_0) \cap B_0(C\epsilon_n)$  for all fixed  $C > 0$ .
- *Assumption 8* (Prior on models) The model prior  $\pi(\mathcal{M})$  satisfies:
  - (i) For any  $\mathcal{M} \supseteq \mathcal{M}_0$ ,  $\frac{\pi(\mathcal{M})}{\pi(\mathcal{M}_0)} \leq r_1$  for some constant  $r_1 > 0$ .
  - (ii) For any  $\mathcal{M}$  such that  $\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset$ ,  $\frac{\pi(\mathcal{M})}{\pi(\mathcal{M}_0)} \leq e^{r_2 p \log n}$  for some constant  $r_2 > 0$ .

Our assumptions on the priors are weak and encompass most of the commonly used priors. Assumption 7 for the priors on the parameter  $\pi(\theta|\mathcal{M})$  are satisfied by, for example, a uniform prior on the model space  $\Theta(\mathcal{M})$ , or a truncated multivariate normal prior on  $\Theta(\mathcal{M})$ . Although we restrict our attention to priors concentrated on the compact set  $\Theta$  whose  $L_2$  norm is bounded by  $R$ , our results can be generalized to priors defined on the whole  $\mathbb{R}^p$ , as long as the tail of  $\pi(\theta|\mathcal{M})$  is sufficiently thin uniformly for all models  $\mathcal{M}$ . Assumption 8 requires that the models larger than the true model  $\mathcal{M}_0$  do not receive overly large prior mass, and the prior on the true model cannot be exponentially small compared to any other models. These requirements can be satisfied by, for example, the prior where each coordinate enters the model independently with a fixed probability, which includes the uniform prior as a special case if this probability is 0.5.

## 2.2 Oracle Properties of BGMM

With all these assumptions, we now state the main results as follows. The proof of Theorem 1 is given in the appendix.

**Theorem 1.** *Suppose Assumptions 1-8 hold. Then*

(i) (Model Selection Consistency)

$$q(\mathcal{M}_0|\mathbf{D}) \rightarrow 1, \text{ w.p.a.1}$$

that is, the quasi-posterior probability of the true model converges to 1, w.p.a.1.

(ii) (Posterior Asymptotic Normality) Given a model  $\mathcal{M}$ , let  $\mathbf{G}_{\mathcal{M}}$  be the submatrix of the derivative matrix  $\mathbf{G}$  with respect to a subvector  $\theta_1$  for  $\theta = (\theta_1^\top, \theta_2^\top)^\top$  in the model  $\mathcal{M}$ . Let  $\bar{\theta}_{\mathcal{M}_0,1} = \theta_{0,\mathcal{M}_0,1} - (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1} \mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta_0)$ , where  $\theta_{0,\mathcal{M}_0,1}$  is the subvector of  $\theta_0$  restricted to  $\Theta(\mathcal{M}_0)$ . Then

$$\sup_{A \subseteq \Theta} \left| \int_A q(\theta|\mathbf{D}) d\theta - \int_{A \cap \Theta(\mathcal{M}_0)} \phi(\theta_1; \bar{\theta}_{\mathcal{M}_0,1}, (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1}/n) \right| \rightarrow 0, \text{ w.p.a.1}$$

where  $\phi(\cdot; \mu, \Sigma)$  is the normal density with mean  $\mu$  and covariance matrix  $\Sigma$ .

Part (i) of Theorem 1 establishes the global model selection consistency of BGMM, similar to previous Bayesian results from Johnson and Rossell (2012) and Liang et al. (2013) for Gaussian regression and generalized linear models. Based on the BGMM posterior, the zero components of the true parameter  $\theta_0$  are estimated to be zero w.p.a.1. This indicates the *superefficiency* of the BGMM method if the true value  $\theta_0$  is sparse as  $p$  grows with  $n$ . It also implies that the MAP model  $\hat{\mathcal{M}}$  will be asymptotically the same as the true model  $\mathcal{M}_0$ . This parallels the frequentist model selection results via penalization for moment based models and estimating equations (Wang et al. 2012, Leng and Tang 2012, Cho and Qu 2013, etc.)

Part (ii) of Theorem 1 establishes an asymptotic normality result, in the sense that the total variation difference between the BGMM posterior measure and a  $k_0$ -dimensional normal distribution concentrated on the true model converges to zero in probability as the sample size increases. This is a direct extension of the Bayesian CLT result in Chernozhukov and Hong (2003) and Belloni and Chernozhukov (2009) from a single full model space to the joint of all submodel spaces. Because the BGMM posterior is a mixture distribution on  $2^p$  model spaces  $\Theta(\mathcal{M})$  with different dimensions, we do not present the total variation distance in terms of the difference between two densities  $q(\theta|\mathbf{D})$  and  $\phi(\theta_1; \bar{\theta}_{\mathcal{M}_0,1}, (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1}/n)$ . The asymptotic mean of the normal distribution  $\bar{\theta}_{\mathcal{M}_0,1}$  is the first order approximation to  $\hat{\theta}_{\mathcal{M}_0,1} = \arg \min_{\theta \in \Theta(\mathcal{M}_0)} \bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta)$ , i.e. the GMM estimator restricted to the subspace  $\Theta(\mathcal{M}_0)$ . Furthermore, given Assumption

6, the generalized information equality is satisfied (Chernozhukov and Hong 2003), and the asymptotic variance of the limit normal distribution is the same as the corresponding frequentist variance of the GMM estimator  $\hat{\theta}_{\mathcal{M}_0,1}$ .

**Remark 1.** *Bayesian oracle property.* The conclusion of Theorem 1 can be written heuristically as

- (i)  $\theta_{0,2}|\mathbf{D} \approx 0$  w.p.a.1;
- (ii)  $q(\theta_{0,1}|\mathbf{D}) \approx N(\bar{\theta}_{\mathcal{M}_0,1}, (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1}/n)$ , w.p.a.1.

The zero components in  $\theta_0$  are estimated to be zero given the data with large probability, and the nonzero components in  $\theta_0$  almost follow a normal distribution centered at the first order approximation to the GMM estimator under the model  $\mathcal{M}_0$ , with the same optimal GMM asymptotic variance matrix, as if the true model  $\mathcal{M}_0$  were known. We call this the *Bayesian oracle property* for model selection, which resembles the frequentist oracle property for penalized likelihood in Fan and Li (2001). Theorem 1 guarantees that the BGMM posterior will automatically identify the unknown true model, and automatically converges to an asymptotic normal distribution centered around the unknown true parameter with the optimal GMM variance, as if the true model were known.

**Remark 2.** *Main idea of the proof.* Our general idea to prove Theorem 1 is to separate all the models  $\mathcal{M}$  different from the true model  $\mathcal{M}_0$  into two groups: (i) the models that miss at least one component of  $\mathcal{M}_0$ , and (ii) the models that contain  $\mathcal{M}_0$  as a strict subset. The selection of a model from the group (i) or (ii) will incur a type I or type II error, respectively. For the models in the first group, we can show that the sum of all posterior odds with respect to  $\mathcal{M}_0$  is upper bounded by  $e^{-Cn \min_{j \in \mathcal{M}_0} |\beta_{0,j}|^2}$  in probability for some constant  $C > 0$ , which goes to zero by the beta-min condition in Assumption 3. For the models in the second group, our proof mainly relies on a quadratic approximation of  $q(\mathbf{D}|\mathcal{M})$  (see Lemma A.2 and Lemma A.3 in the appendix). For a heuristic argument, when  $\mathcal{M} \supseteq \mathcal{M}_0$ , the BGMM likelihood can be approximated by an unnormalized  $|\mathcal{M}|$ -dimensional normal density centered at  $\bar{\theta}_{\mathcal{M},1}$ , the first order approximation of the GMM estimator on  $\Theta(\mathcal{M})$ . With constants and smaller order terms ignored, written in the logarithm form, this is

$$\begin{aligned} \log q(\mathbf{D}|\theta, \mathcal{M}) &\approx -\frac{n}{2}(\theta_1 - \bar{\theta}_{\mathcal{M},1})^\top \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}}(\theta_1 - \bar{\theta}_{\mathcal{M},1}) \\ &\approx \log \phi(\theta_1; \bar{\theta}_{\mathcal{M},1}, (\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})^{-1}/n) - \frac{|\mathcal{M}| \log n}{2}, \end{aligned}$$



where  $\phi(\cdot)$  denotes the normal density. Therefore, the BGMM is intrinsically equipped with a BIC-type penalty that leads to consistent model selection. The frequentist BIC-based model selection for models with unconditional moment restrictions has been studied by Andrews (1999), Andrews and Lu (2001), etc., under a fixed dimension  $p$ . Our result shows that such a BIC-penalty allows consistent model selection as long as the growth rate of  $p$  satisfies Assumption 2.

**Remark 3.** In Theorem 1 we presented an asymptotic normal distribution centered at the first order approximation  $\bar{\theta}_{\mathcal{M}_0,1}$ . This is in the same line with Belloni and Chernozhukov (2009) and it applies to all our motivating examples. We can further replace this approximation  $\bar{\theta}_{\mathcal{M}_0,1}$  with the exact GMM estimator  $\hat{\theta}_{\mathcal{M}_0,1}$  on the space  $\Theta(\mathcal{M}_0)$ , using the following high level assumption.

- *Assumption 9* (High Order Approximation) For the true model  $\mathcal{M}_0$ ,  $\|\bar{\theta}_{\mathcal{M}_0,1} - \hat{\theta}_{\mathcal{M}_0,1}\| = O_p(p^{\frac{3}{2}}/n)$ .

Then we have the following corollary. The proof is given in the Appendix.

**Corollary 1.** (*Posterior Asymptotic Normality*) Suppose Assumptions 1-9 hold. Given a model  $\mathcal{M}$ , let  $\mathbf{G}_{\mathcal{M}}$  be the submatrix of the derivative matrix  $\mathbf{G}$  with respect to a subvector  $\theta_1$  for  $\theta = (\theta_1^\top, \theta_2^\top)^\top$  in the model  $\mathcal{M}$ . Let  $\hat{\theta}_{\mathcal{M}_0,1} = \arg \min_{\theta \in \Theta(\mathcal{M}_0)} \bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta)$ . Then

$$\sup_{A \subseteq \Theta} \left| \int_A q(\theta | \mathbf{D}) d\theta - \int_{A \cap \Theta(\mathcal{M}_0)} \phi(\theta_1; \hat{\theta}_{\mathcal{M}_0,1}, (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1}/n) \right| \rightarrow 0, \text{ w.p.a.1}$$

where  $\phi(\cdot; \mu, \Sigma)$  is the normal density with mean  $\mu$  and covariance matrix  $\Sigma$ .

Now the BGMM posterior will asymptotically center at the exact GMM estimator  $\hat{\theta}_{\mathcal{M}_0,1}$ , at the cost of the extra Assumption 9, which is a high level assumption that involves more regularity conditions on the moment  $g(D, \theta)$ . The difference between  $\hat{\theta}_{\mathcal{M}_0,1}$  and  $\bar{\theta}_{\mathcal{M}_0,1}$  is the asymptotic higher order bias of the GMM estimator. Therefore, when  $p$  does not increase with  $n$ , one can directly apply the set of conditions in Newey and Smith (2004) to check Assumption 9. When  $p$  increases with  $n$  but  $k_0 = O(1)$ , i.e. the true model is sparse, we can adapt the assumptions of Newey and Smith (2004) as follows:  $E \sup_{\Theta} \|g(D, \theta)/\sqrt{p}\|^{2+\delta} < \infty$  for some  $\delta > 0$ ;  $g(D, \theta)$  is four times differentiable; there exists a neighborhood  $B_0(\eta) \subset \Theta(\mathcal{M}_0)$  of  $\theta_0$  for some  $\eta > 0$ , and a function  $b(D)$  of the random vector  $D$  with  $E[b(D)^6] < \infty$ , such that  $\sup_{B_0(\eta)} \|\nabla^j g(D, \theta)/\sqrt{p}\| \leq b(D)$  for

$j = 0, 1, \dots, 4$ , and  $\|\nabla^j g(D, \theta) - \nabla^j g(D, \theta_0)\|/\sqrt{p} \leq b(D)\|\theta - \theta_0\|$  for all  $\theta \in B_0(\eta)$  and  $j = 0, 1, \dots, 4$ ;  $\mathbf{V}_n = \mathbf{V} + \sum_{i=1}^n \xi(D_i)/n + O_p(p/n)$  with  $E[\xi(D)] = 0$  and  $E[\|\xi(D)/\sqrt{p}\|^6] < \infty$ . Because the norm of the derivative  $\nabla^j g(D, \theta)/\sqrt{p}$  only scales with  $k_0$  but not  $p$  if  $\theta \in \Theta(\mathcal{M}_0)$ , one can verify that the asymptotic higher order bias of  $\hat{\theta}_{\mathcal{M}_0,1}$  is at most  $O_p(p^{3/2}/n)$  for all the four bias components in the stochastic expansion of the GMM estimator in Theorem 4.1 of Newey and Smith (2004). Note that these extra assumptions include the differentiability of the moment  $g(D, \theta)$ , which holds for our correlated longitudinal data example and the graphical partial correlation selection example, but excludes the application of Assumption 9 and Corollary 1 to the quantile regression example.

## 2.3 Application to Motivating Examples

### 2.3.1 Correlated Longitudinal Data

We use the same notations as in the introduction. Without loss of generality, we assume  $Y_i$  and each  $X_{ijk}$  has been centered such that  $EY_i = 0$  and  $EX_{ijk} = 0$ , for  $i = 1, \dots, n$ ,  $j = 1, \dots, s$  and  $k = 1, \dots, p$ . For the ease of presentation and the simplification of our proofs, we assume that the working correlation matrix  $\mathbf{R}$  is correctly specified and does not depend on  $\theta$ . We also plug in a preliminary consistent estimator  $\tilde{\theta}$  for  $\theta$  to the nonlinear part  $\frac{\partial \mu_i(\theta)^\top}{\partial \theta} \mathbf{S}_i^{-1}$  of the moment condition  $g(D_i, \theta)$ . Such consistent estimator  $\tilde{\theta}$  exists even with growing  $p$ . For example, one can take  $\tilde{\theta}$  to be the solution of estimating equations  $\sum_{i=1}^n \mathbf{X}_i(Y_i - \mu_i(\theta)) = 0$ . Under the assumptions given in the theorem below, one can show that  $\|\tilde{\theta} - \theta_0\| = O_p(\sqrt{p/n})$  similar to Example 1 in Wang (2011).

The matrix  $\mathbf{V}_n$  in BGMM can be taken as  $\mathbf{V}_n = n^{-1} \sum_{i=1}^n (g(D_i, \tilde{\theta}) - \bar{g}(\mathbf{D}, \tilde{\theta}))(g(D_i, \tilde{\theta}) - \bar{g}(\mathbf{D}, \tilde{\theta}))^\top$ , where  $\tilde{\theta}$  is any consistent preliminary estimator of  $\theta_0$ . Given the growth rate of  $p$  in Assumption 2, one can show that  $\mathbf{V}_n$  converges in operator norm to  $\mathbf{V} = \text{Var}(g(D, \theta_0))$ .

Let  $\dot{\mu}(x)$  and  $\ddot{\mu}(x)$  be the first and the second derivatives of  $\mu(x)$ . We then have the following theorem for the BGMM based on the moment condition (5) for the correlated longitudinal data.

**Theorem 2.** *For the moment (5), suppose that Assumptions 1, 2, 3, 7 and 8 hold, and in Assumption 2 let  $\alpha = 1$ . Suppose  $\tilde{\theta}$  is the preliminary estimator that solves  $\sum_{i=1}^n \mathbf{X}_i(Y_i - \mu_i(\theta)) = 0$  and  $\mathbf{V}_n = n^{-1} \sum_{i=1}^n (g(D_i, \tilde{\theta}) - \bar{g}(\mathbf{D}, \tilde{\theta}))(g(D_i, \tilde{\theta}) - \bar{g}(\mathbf{D}, \tilde{\theta}))^\top$ .*

In addition, if

(1)  $|X_{ijk}| \leq C_X$  almost surely for some large constant  $C_X > 0$  and all  $i = 1, \dots, n$ ,  $j = 1, \dots, s$ ,  $k = 1, \dots, p$ .  $\sup_{1 \leq j \leq s} E(Y_j^4) < \infty$ ;

(2)  $E[\mathbf{X}_i^\top \mathbf{X}_i]$  and  $\mathbf{R}$  have eigenvalues bounded above and below by constants w.p.a.1 for all  $i = 1, \dots, n$ ;

(3)  $\dot{\mu}(X_{ij}^\top \theta)$  is bounded above and below uniformly for all possible values of  $X_{ij}$  and  $\theta \in \Theta$ , w.p.a.1.  $\ddot{\mu}(X_{ij}^\top \theta)$  is bounded above, and  $\phi(X_{ij}^\top \theta)$  is bounded above and below uniformly for all  $X_{ij}$  and  $\theta \in B_0(c\epsilon_n)$  w.p.a.1 for any fixed  $c > 0$ ;

then Assumptions 4, 5 and 6 also hold. Therefore BGMM for (5) satisfies the Bayesian oracle property in Theorem 1.

**Remark 4.** In condition (1) we impose an absolute bound on all the covariates for convenience, though this can be replaced by relaxed conditions on the tail behavior or the high order moments on  $X_{ijk}$ . Condition (2) for eigenvalues are standard. Here for simplicity, we use only one working correlation matrix  $\mathbf{R}$  such that  $m = p$ , though the result can be easily extended to more than one working correlation matrices like in Qu et al. (2000). Condition (3) requires certain bounds on the derivatives of  $\mu$  and also  $\phi$ . In particular,  $\mu(t) = t$  for linear regression trivially satisfies this condition. For logistic regression,  $\mu(t) = e^t/(1 + e^t)$  and  $\phi(t) = e^t/(1 + e^t)^2$ . If we only consider the models with  $O(1)$  size, or restrict the  $L_2$  norm of  $X$  by a large constant, then condition (3) is also satisfied for the derivatives of  $\mu$  and  $\phi$  evaluated at  $X_{ij}^\top \theta$ . Similar arguments can be applied to Poisson regression, exponential regression and probit regression, etc. In Liang and Zeger (1986) and Wang et al. (2012), the marginal density of  $Y_{ij}$  is modeled as a canonical exponential family, with  $\text{Var}(Y_{ij}|X_{ij}) = \psi \dot{\mu}(X_{ij}^\top \theta)$ , where  $\psi$  is the dispersion parameter. Here we have considered a general form of the function  $\phi$  and therefore our setup includes theirs as a special case.

### 2.3.2 Quantile Regression

Without loss of generality, we assume that the random variables  $Y$  and  $X$  are centered such that  $EY = 0$  and  $EX = 0$ . The conditional distribution  $F_{Y|X}$  is assumed to be continuous, and let  $f_{Y|X}$  be its conditional density. It can be calculated that  $\mathbf{V} = \text{Var}(g(D, \theta_0)) = \tau(1 - \tau)E[XX^\top]$  for the unconditional moments (6), and we can estimate  $\mathbf{V}$  by  $\mathbf{V}_n = n^{-1}\tau(1 - \tau)\sum_{i=1}^n X_i X_i^\top$ . Then we have the following theorem about quantile regression.

**Theorem 3.** *For the moment (6), suppose that Assumptions 1, 2, 3, 7 and 8 hold, and in Assumption 2 let  $\alpha = 1/2$ . Suppose that  $\mathbf{V}_n = n^{-1}\tau(1 - \tau) \sum_{i=1}^n X_i X_i^\top$ . In addition, if*

- (1) *For any generic random vector  $X = (X_1, \dots, X_p)^\top$ ,  $|X_j| \leq C_X$  almost surely for some large constants  $C_X > 0$  and all  $j = 1, \dots, p$ ;*
- (2)  *$f_{Y|X}$  is continuously differentiable with the first derivative  $\dot{f}_{Y|X}$ .  $f_{Y|X}$  and  $\dot{f}_{Y|X}$  are almost surely bounded above on the support of  $Y$  for any value of  $X$ .  $f_{Y|X}$  is further bounded below for any value of  $X$ .*
- (3)  *$E[XX^\top]$  has eigenvalues bounded above and below by constants.*

*then Assumptions 4, 5 and 6 also hold. The BGMM for (6) satisfies the Bayesian oracle property in Theorem 1.*

**Remark 5.** The quantile regression example can be generalized to the instrumental variable quantile regression model (IVQR), as discussed in Chernozhukov and Hansen (2005, 2006). In the IVQR, the predictor  $X$  could contain endogenous components, and we can still consistently estimate the parameter  $\theta$  using other informative and exogenous instrumental variables. The model formulation will be more complicated but can be incorporated into the BGMM framework using the unconditional moments based on IV (e.g. Chernozhukov and Hong 2003).

### 2.3.3 Partial Correlation Selection

We use the same notation as in the introduction. Suppose the true covariance matrix is  $\Sigma_0$  and the true precision matrix is  $\Omega_0$ , whose dimensions are  $s \times s$ . Each parameter  $\theta$  we consider here corresponds to a positive definite matrix  $\Omega$ , since  $\theta$  comes from the vectorization of the upper triangle of such a  $\Omega$ . The true parameter  $\theta_0$  comes from  $\Omega_0$ . We can take  $\mathbf{V}_n = n^{-1} \sum_{i=1}^n (g(D_i, \tilde{\theta}) - \bar{g}(\mathbf{D}, \tilde{\theta}))(g(D_i, \tilde{\theta}) - \bar{g}(\mathbf{D}, \tilde{\theta}))^\top$ , where  $\tilde{\theta}$  is the estimated parameter by inverting the empirical covariance matrix. Then we have the following theorem for partial correlation selection.

**Theorem 4.** *For the moment (7), suppose that Assumptions 1, 2, 3, 7 and 8 hold for  $p = s(s + 1)/2$ , and in Assumption 2 let  $\alpha = 1$ . In addition, if*

- (1) *Uniformly for all  $\theta \in \Theta$ , the corresponding  $\Omega$  has eigenvalues bounded above and below by constants;*
- (2) *For any random vector  $Y = (Y_1, \dots, Y_s)^\top$ ,  $\sup_{1 \leq j \leq s} E(Y_j^8) < \infty$ ;*
- (3) *The eigenvalues of  $\mathbf{V} = \text{Var}[g(D, \theta_0)]$  are bounded from above and below by constant;*

then Assumptions 4, 5 and 6 also hold. Therefore BGMM (7) satisfies the Bayesian oracle property in Theorem 1.

**Remark 6.** Here we restrict the space of the precision matrix  $\mathbf{\Omega}$  to a (possibly large) convex and compact set, with boundaries set by the smallest and the largest eigenvalues of  $\mathbf{\Omega}$  as in the condition (1). The boundedness of  $\sup_j E(Y_j^8)$  in the condition (2) is to guarantee the convergence of  $\mathbf{V}_n$  to  $\mathbf{V}$ . Here we have directly assumed that the eigenvalues of  $\mathbf{V}$  are bounded from above and below, mainly because this condition is not trivial and can hardly be obtained from any low level conditions. In fact, there could be cases where  $Y$  has a nondegenerate covariance matrix  $\mathbf{\Sigma}$ , but  $\mathbf{V}$  is degenerate. For example, let  $s \geq 4$ ,  $Z \sim \text{Uniform}[1, 2]$  independent of  $Y$ , and  $Y = (Y_1, \dots, Y_s)$  where  $(Y_1, Y_2)$  are independent  $N(0, 1)$ ,  $Y_3 = Y_1 \cdot Z$ ,  $Y_4 = Y_2/Z$ ,  $(Y_5, \dots, Y_s)$  are independent  $N(0, 1)$  and are independent of  $Y_1, Y_2, Z$ . Then it can be verified that  $\text{Cov}(Y_1, Y_3) = 3/2$ ,  $\text{Cov}(Y_2, Y_4) = \log(2)$  and all other pairwise covariances are zeros. Hence  $\mathbf{\Sigma}$  is nondegenerate, but since  $Y_1 Y_2 \equiv Y_3 Y_4$ ,  $\mathbf{V} = \text{Var}[g(D, \theta_0)]$  is degenerate because two components of  $g(D, \theta)$  are exactly identical. Therefore in practice, Condition (3) may or may not hold depending on the distribution of  $Y$ .

## 2.4 Asymptotic Validity of the BGMM Posterior

As shown in Kim (2002), the limited information likelihood we have used for BGMM provides a large sample approximation to the true likelihood function of  $\theta$  given the moment restrictions  $Eg(D, \theta) = 0$ . One may ask about how well this approximation could be. For the validity of usual Bayesian inference, such as constructing the Bayesian credible sets, it is necessary and sufficient to impose Assumption 6 that  $\mathbf{V}_n$  consistently estimates  $\mathbf{V}$ , i.e.  $\mathbf{V}_n$  satisfies the generalized information equality as in Kim (2002) and Chernozhukov and Hong (2003). However, due to the limited information contained in  $Eg(D, \theta) = 0$ , in general one cannot expect the LIL  $q(\mathbf{D}|\theta)$  to coincide with the true likelihood function  $p(\mathbf{D}|\theta)$ . Instead the quasi-posterior  $q(\theta|\mathbf{D})$  can be used to approximate the posterior of  $\theta$  given some summary statistic from the sample. Let  $\hat{\theta}$  be the minimizer of the GMM criterion function  $\bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}^{-1} \bar{g}(\mathbf{D}, \theta)$  over the full  $p$ -dimensional model space. So  $\hat{\theta}$  is implicitly a statistic of the sample  $\mathbf{D}$ , and it does not depend on  $\theta_0$  and the unknown true model  $\mathcal{M}_0$ . Since from Corollary 1, the BGMM posterior is asymptotically centered at the GMM estimator, one could conjecture that the LIL  $q(\mathbf{D}|\theta)$  approximates the density  $p(\hat{\theta}|\theta)$  of  $\hat{\theta}$ . Accordingly, the BGMM posterior  $q(\theta|\mathbf{D})$  approximates the pos-

terior  $p(\theta|\hat{\theta})$  of  $\theta$  given  $\hat{\theta}$ , at least in the first order asymptotics. In the following, we formalize this idea and show more general results under the model selection setup.

For two generic models  $M_1$  and  $M_2$ , we define the Bayes factor based on  $p(\hat{\theta}|\theta)$  as

$$\text{BF}_{\hat{\theta}}[\mathcal{M}_1 : \mathcal{M}_2] = \frac{p(\hat{\theta}|\mathcal{M}_1)}{p(\hat{\theta}|\mathcal{M}_2)} = \frac{\int_{\Theta(\mathcal{M}_1)} p(\hat{\theta}|\theta)\pi(\theta|\mathcal{M}_1)d\theta}{\int_{\Theta(\mathcal{M}_2)} p(\hat{\theta}|\theta)\pi(\theta|\mathcal{M}_2)d\theta}$$

For the ease of presentation, we focus on the situation with a nonincreasing dimension  $p$ , and make the following additional assumption.

- (*Assumption 10*) (i)  $\dim(\theta) = p$  and  $1 \leq p \leq \bar{p}$ , for some large fixed integer  $\bar{p}$ .
- (ii)  $\min_{j \in \mathcal{M}_0} |\theta_{0,j}| \geq \underline{\theta}$  for some small constant  $\underline{\theta} > 0$ .
- (iii) Let  $\mathbf{V}(\theta) = \text{Var}[g(D, \theta)]$  and  $\mathbf{G}(\theta) = \nabla_{\theta} E g(D, \theta)$ . Then the elements of  $\mathbf{V}(\theta)$  and  $\mathbf{G}(\theta)$  are continuous functions of  $\theta$ , and the eigenvalues of  $\mathbf{G}(\theta)^{\top} \mathbf{G}(\theta)$  and  $\mathbf{V}(\theta)$  are uniformly bounded below and above for all  $\theta \in \Theta$ .
- (iv) For any two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , there exists a constant  $r > 0$  such that  $\frac{\pi(\mathcal{M}_2)}{\pi(\mathcal{M}_1)} \leq r$ .
- (v)  $\|\hat{\theta} - \bar{\theta}\| = O_p(1/n)$ , where  $\hat{\theta}$  is the GMM estimator on the full model space, and  $\bar{\theta} = \theta_0 - (\mathbf{G}^{\top} \mathbf{V}^{-1} \mathbf{G})^{-1} \mathbf{G}^{\top} \mathbf{V}^{-1} \bar{g}(\mathbf{D}, \theta_0)$ .

The strengthened beta-min condition in (ii) is to emphasize the difference between the models that make the type I error and the type II error. According to theorems we are going to present below, the models in the former group have an exponentially small  $\text{BF}_q[\mathcal{M} : \mathcal{M}_0]$ , while the models in the latter group have a polynomially small  $\text{BF}_q[\mathcal{M} : \mathcal{M}_0]$ . This is also the essential behavior from the Bayesian hypothesis test, which favors the true alternative hypothesis more, as discovered in Johnson and Rossell (2010). We will show that similar behavior is also shared by  $\text{BF}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0]$ , and hereby establish a correspondence between the BGMM method and the exact Bayesian method given  $\hat{\theta}$ .

Part (iii) assumes the continuity of the matrices in  $\theta$  and also the uniform bound for eigenvalues. This is a mild assumption given the compactness of  $\Theta$ . Part (iv) has strengthened Assumption 8 and required that no model should be assigned extremely large or small prior. Part (v) is an extended version of Assumption 9 to the GMM estimator on the full model space. Since we are now considering bounded dimension  $p$ , this assumption will hold trivially under similar conditions to Newey and Smith (2004).

Let  $\mathbf{F}(\theta)$  be a  $p \times p$  matrix such that  $\mathbf{F}(\theta)^\top \mathbf{F}(\theta) = \mathbf{G}(\theta)^\top \mathbf{V}(\theta)^{-1} \mathbf{G}(\theta)$ . Define  $Z = \sqrt{n} \mathbf{F}(\theta)(\hat{\theta} - \theta)$ . Then  $Z$  is asymptotically  $p$ -dimensional standard normal if the true parameter is  $\theta_0 = \theta$ . We impose the following high level assumption on the difference between the exact density function  $p_Z(z)$  of  $Z$  and the normal density.

- *Assumption 11* (Uniform Bound) As  $n \rightarrow \infty$ ,

$$\sup_{\theta \in \Theta} \sup_z (1 + \|z\|^{p+1}) |p_Z(z|\theta) - \phi(z; 0, \mathbf{I}_p)| = \tau_n,$$

where  $\tau_n = o(1)$  does not depend on  $z$  and  $\theta$ , and  $\mathbf{I}_p$  is the  $p \times p$  identity matrix.

Assumption 11 claims that the difference between the density of the normalized GMM estimator  $Z$  and its asymptotic limit of normal density can be uniformly bounded by an integrable function  $c(\|z\|) = 1/(1 + \|z\|^{p+1})$ , and the uniformity is for both the value of  $z$  and the parameter  $\theta$  in the compact space  $\Theta$ . This is a high level condition that originates from the Condition E in Yuan and Clarke (2004). We do not intend to give a full proof of it under low level assumptions, but we explain why it is a reasonable assumption below.

Consider the case where the  $(p + 1)$ -th moment of  $g(D, \theta)$  exists. To show Assumption 11, we proceed in several steps. First, under similar regularity conditions that make Assumption 10(v) hold, one can see that for a fixed  $\theta$ , the density of  $Z$  is asymptotically uniformly close to the density of the normalized first order approximation  $\bar{Z} = \sqrt{n} \mathbf{F}(\theta)(\bar{\theta} - \theta)$ , up to the order  $O(1/\sqrt{n})$ , where  $\bar{\theta}$  is defined in Theorem 1 (ii). See Kundhi and Rilstone (2012, 2013) for the formal proofs of a general class of nonlinear estimators, which can also be applied to the GMM estimator. Second, due to the sample average form of  $\bar{\theta}$  and hence  $\bar{Z}$ , one can use Proposition 1 in Yuan and Clarke (2004) and take  $c(x) = 1/(1 + x^{p+1})$ . This proposition provides a bound for the difference between the density of  $\bar{Z}$  and its limiting normal density, which holds uniformly for all  $\theta \in \Theta$ . Its proof involves the techniques in Chap.19 of Bhattacharya and Ranga Rao (1986) about the uniform convergence of continuous characteristic functions in the compact set  $\Theta$ . Third, one can show that in Proposition 1 of Yuan and Clarke (2004), the summation of the Edgeworth series beyond the leading normal density term has the order  $o_p(1)$ . This is due to the finite moments of  $g(D, \theta)$  up to the  $(p + 1)$ -th order, as well as the boundness of the multivariate Hermite polynomials. Finally we combine all these pieces and conclude that the uniform deviation in Assumption 11 holds with some  $\tau_n = o(1)$ .

The next theorem provides a comparison between the convergence rates for the Bayes Factors  $\text{BF}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0]$  from the likelihood given the statistic  $\hat{\theta}$  with  $\text{BF}_q[\mathcal{M} : \mathcal{M}_0]$  from the BGMM method.

**Theorem 5.** (*Equivalence of Bayes Factors*) Suppose Assumptions 1-11 hold, and the true model size is  $|\mathcal{M}_0| = k_0$ . Then under the same prior  $\pi(\theta|\mathcal{M})$  and  $\pi(\mathcal{M})$ , as  $n \rightarrow \infty$  w.p.a.1,

(i) For any model  $\mathcal{M}$  with  $\mathcal{M} \supseteq \mathcal{M}_0$ ,

$$\begin{aligned} \frac{\text{BF}_q[\mathcal{M} : \mathcal{M}_0]}{\text{BF}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0]} &\rightarrow 1; \\ \text{BF}_q[\mathcal{M} : \mathcal{M}_0] &\asymp \text{BF}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] \asymp n^{-\frac{|\mathcal{M}| - k_0}{2}} \succeq n^{-\frac{p - k_0}{2}}; \end{aligned}$$

(ii) For any model with  $\mathcal{M}$  with  $\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset$ , there exists a constant  $C > 0$ , such that

$$\begin{aligned} \text{BF}_q[\mathcal{M} : \mathcal{M}_0] &\leq \exp(-Cn\underline{\theta}^2) \prec n^{-\frac{p - k_0 + 1}{2}}; \\ \text{BF}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] &\leq \exp(-Cn\underline{\theta}^2) \vee \tau_n n^{-\frac{p - k_0 + 1}{2}} \prec n^{-\frac{p - k_0 + 1}{2}}. \end{aligned}$$

Theorem 5 compares the Bayes factors from BGMM and  $p(\hat{\theta}|\theta)$ , for the models that make a type I error (Part ii) and a type II error (Part i). The theorem has at least two direct implications. First, for the models that make a type II error (including more components of  $\theta$  than necessary), the Bayes factors are asymptotically equal, and both decrease polynomially in the sample size  $n$ . The polynomial index reflects the difference in dimensions between  $\mathcal{M}$  and  $\mathcal{M}_0$ . Second, for the models make a type I error (missing at least one nonzero component in  $\theta_0$ ), the Bayes factor from BGMM decreases exponentially fast in  $n$ . For the Bayes factor from  $p(\hat{\theta}|\theta)$ , we have obtained an upper bound for its rate, which also depends on the rate  $\tau_n$  in Assumption 11 besides the usual exponential rate. Because  $\tau_n = o(1)$  by Assumption 11, we can see clearly that there exists at least a  $n^{-1/2}$  gap between the convergence rates of Bayes factors for the models with type I and type II errors, with the threshold rate  $n^{-\frac{p - k_0 + 1}{2}}$ , which depends on the unknown dimension  $k_0$  of the true model  $\mathcal{M}_0$ . In general, the posterior probabilities of the models with the type I error converge faster to zero than the posterior of the models with the type II error.

This extra part  $\tau_n n^{-\frac{p - k_0 + 1}{2}}$  for the Bayes factor in (ii) arises mainly technically from our Assumption 11. Usually,  $\tau_n = o(1)$  in Assumption 11 is tight and may not be improved. However, we conjecture that it could be removed by making stronger assumptions on the density function  $p(\hat{\theta}|\theta)$ , or the normalized density  $p_Z(z|\theta)$ . For example,



$p_Z(\sqrt{n}\mathbf{F}(\theta)(\hat{\theta} - \theta)|\theta)$  decreases exponentially fast in  $n$  as  $\theta$  moves away from the true parameter  $\theta_0$ . However, we realized that usually it is difficult to check such assumptions for the distribution of  $\hat{\theta}$  because it does not have an explicit density, except for a few special cases where  $\hat{\theta}$  comes from the exponential family. We also note that such compromised rate also shows up in Marin et al. (2013) Lemma 1, where they studied the convergence rates of the Bayes factor given a general statistic. Although typically one cannot obtain the exact form of the density  $p(\hat{\theta}|\theta)$  and its posterior  $p(\theta|\hat{\theta})$ , Theorem 5 provides some evidence that in the asymptotic sense, the Bayes factors from the BGMM method behave very similarly to the Bayes factor from  $p(\hat{\theta}|\theta)$ , indicating the validity of using  $\text{BF}_q[\mathcal{M} : \mathcal{M}_0]$  for model selection.

**Remark 7.** In principle, Theorem 5 provides a guideline to interpret the BGMM posterior probabilities of different models. For simplicity, suppose that all models receive the uniform prior  $\pi(\mathcal{M}) \propto 1$ . Then since  $q(\mathcal{M}_0|\mathbf{D}) \rightarrow 1$ , the posterior  $q(\mathcal{M}|\mathbf{D})$  is roughly the same as  $\text{BF}_q[\mathcal{M} : \mathcal{M}_0]$ . Because of the gap between the polynomial rate in (i) and the exponential rate in (ii), we can choose any rate in between as a threshold, for example  $e^{-\sqrt{n}}$ . If a model  $\mathcal{M}$  has  $q(\mathcal{M}|\mathbf{D}) \geq e^{-\sqrt{n}}$ , then we can approximately regard  $q(\mathcal{M}|\mathbf{D})$  as the true posterior probability  $p(\mathcal{M}|\hat{\theta})$  and consider  $\mathcal{M}$  as a model with nonnegligible posterior. This fits well with the common practice that we rank the models according to their posterior probabilities and only study the models on top of the list.

Based on Theorem 5, we can further show that the BGMM posterior  $q(\theta|\mathbf{D})$  and the exact posterior  $p(\theta|\hat{\theta})$  are close in the total variation distance asymptotically.

**Theorem 6.** *Suppose Assumptions 1-11 hold. Let the full model be  $\mathcal{M}_{\text{full}}$ . Then under the same prior  $\pi(\theta|\mathcal{M})$  and  $\pi(\mathcal{M})$ , as  $n \rightarrow \infty$  w.p.a.1,*

*(i) (Model Selection Convergence Rate) If  $\mathcal{M}_0 \neq \mathcal{M}_{\text{full}}$ , then*

$$\begin{aligned} \frac{q(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\mathbf{D})}{p(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\hat{\theta})} &\rightarrow 1; \\ q(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\mathbf{D}) &\asymp p(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\hat{\theta}) \asymp n^{-\frac{1}{2}} \rightarrow 0; \end{aligned}$$

*If  $\mathcal{M}_0 = \mathcal{M}_{\text{full}}$ , then for some constant  $C > 0$ ,*

$$\begin{aligned} q(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\mathbf{D}) &\leq \exp(-Cn\underline{\theta}^2) \rightarrow 0; \\ p(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\hat{\theta}) &\leq \exp(-Cn\underline{\theta}^2) \vee \tau_n n^{-\frac{p-k_0+1}{2}} \rightarrow 0. \end{aligned}$$

(ii) (*Asymptotic Posterior Validity*)

$$\sup_{A \subseteq \Theta} \left| \int_A q(\theta|\mathbf{D})d\theta - \int_A p(\theta|\hat{\theta})d\theta \right| \rightarrow 0.$$

Part (i) of the theorem is a direct corollary from Theorem 5. It implies that the posterior probability of the true model  $\mathcal{M}_0$  converges to 1 at exactly the same rate using either the BGMM or  $p(\hat{\theta}|\theta)$ , when the true model is a strict submodel of the full model. When the true model is exactly the same as the full model, we have only upper bounds for the model selection convergence rates, as they usually decrease exponentially fast, but again the rate is compromised by  $\tau_n$  from Assumption 11 when we consider  $p(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0|\hat{\theta})$ . In either scenario, we have the global model selection consistency for both the BGMM posterior and the posterior given  $\hat{\theta}$ .

Part (ii) gives the asymptotic validity of the BGMM posterior, in the sense that it provides the same asymptotic inference as the exact posterior of  $\theta$  given the statistic  $\hat{\theta}$ . It has the immediate implication that the posterior credible sets for the parameters constructed from the BGMM posterior are asymptotically valid. It is worth noting that the conclusion of (ii) is only related to the global model selection consistency for both the BGMM posterior and the posterior of  $p(\theta|\hat{\theta})$ , and does not depend on the exact convergence rates of model selection in Part (i). In fact, Part (ii) also holds for general non-model selection prior  $\pi(\theta)$  as long as it has a bounded continuous density on  $\Theta$ . This can be obtained from combining Theorem 1 in Chernozhukov and Hong (2003) and Theorem 2 in Yuan and Clarke (2004) (where  $T_n = \hat{\theta}$ ), under Condition E in Yuan and Clarke (2004). Our proof of Theorem 6 follows a similar route by using Assumption 11, but has accommodated the nature of model selection priors  $\pi(\theta, \mathcal{M}) = \pi(\theta|\mathcal{M})\pi(\mathcal{M})$  in Assumption 7 and 8.

**Remark 8.** We have discussed the asymptotic closeness of the BGMM posterior to the posterior given the GMM estimator  $\hat{\theta}$ . One can further explore the higher order asymptotics of  $q(\theta|\mathbf{D})$  and  $p(\theta|\hat{\theta})$ , for example expanding both posterior densities as Edgeworth series of the asymptotic pivotal quantity  $\sqrt{n}\mathbf{F}(\theta)(\theta - \hat{\theta})$ . In this sense, our result in Part (ii) of Theorem 6 only captures the leading order closeness from  $q(\theta|\mathbf{D})$  to  $p(\theta|\hat{\theta})$ . However, we conjecture that in general the higher order terms of  $q(\theta|\mathbf{D})$  and  $p(\theta|\hat{\theta})$  do not match with each other, since the LIL takes a quadratic form of the moment conditions while the true density of  $\hat{\theta}$  depends on other features of  $P_{\mathbf{D}}$ , such as the high

order moments. Similar work in this direction includes Fang and Mukerjee (2006), where they have shown by a simple example of sample mean that the Edgeworth expansions from the empirical likelihood and the density of the sample average do not agree in high order terms.

## 3 Numerical Examples

### 3.1 Algorithm

Because the LIL (2) allows any form of moment conditions  $g(D, \theta)$ , usually one cannot derive an analytical close-form solution for the BGMM posterior update. Therefore, we adopt a reversible jump MCMC algorithm with Metropolis moves both between models and within a model to explore the full posterior of  $\theta$ , similar in spirit to the MCMC algorithm for the Gibbs posterior model selection (Jiang and Tanner 2008, Chen et al. 2010), and also the PAC-Bayesian model selection (Rigollet and Tsybakov 2011, Guedj and Alquier 2013, Alquier and Biau 2013). In the  $i$ th iteration, the between-model steps either add a new component to the nonzero part of  $\theta^{(i)}$ , or remove an existing component in the nonzero part of  $\theta^{(i)}$ , each with probability 0.5. When we add a new component, the parameter value for this new component is sampled from  $N(0, \sigma_{\text{add}}^2)$ , while the values of the existing components in  $\theta^{(i)}$  are retained. Both the “add” and the “remove” operations will be accepted or rejected with a probability based on the ratio of the posteriors evaluated at the new proposed parameter and the current parameter. This between-model step is then followed by a within-model step, in which we draw a new parameter value in the same model as  $\theta^{(i)}$  from a proposal distribution. In practice, to efficiently explore each model space, we use a normal distribution as a proposal distribution, with mean zero and a properly chosen variance  $c \cdot \Xi_M$ . Here  $\Xi_M$  is the submatrix of  $\Xi$  with rows and columns corresponding to the model  $\mathcal{M}$ , and  $\Xi$  is an estimated covariance matrix for the GMM estimator  $\hat{\theta}$ , which can be obtained numerically by inverting the Hessian matrix at the preliminary one-step GMM estimator  $\tilde{\theta}$  on the full model space. We set  $c = 2.4^2$  as suggested in Gelman et al. (2013) to achieve the ideal acceptance rate for within-model Metropolis moves. We also run pilot chains to tune the value of  $\sigma_{\text{add}}$  for better mixing of the Markov chain. As a result, the Markov chain consists of  $\theta^{(i)}$  drawn from the full BGMM posterior across different model spaces.

---

**Algorithm 1** Model Selection Algorithm for Bayesian GMM

---

Set  $\theta^{(0)} = 0$ ,  $\mathcal{M}^{(0)} = \{\text{the null model}\}$ .

**for**  $i = 1$  to  $N$  **do**

**(Between-model Step)**

    Set  $\theta^{\text{new}} = \theta^{(i)}$ .

    Draw  $u_1 \sim \text{Uniform}(0, 1)$ .

**if**  $u_1 \leq 0.5$  **then**

        Choose one of the zero components (suppose it is the  $j$ th component) of  $\theta^{(i)}$  with probability  $1/(p - |\mathcal{M}^{(i)}|)$ .

        Set the  $j$ th component in  $\theta^{\text{new}}$  to  $\theta_j \sim \theta_N(0, \sigma_{\text{add}}^2)$ .

        Set  $\theta^{i+1} = \theta^{\text{new}}$  with probability

$$\alpha_1 = \min \left( \frac{q(\theta^{\text{new}}|\mathbf{D})}{q(\theta^{(i)}|\mathbf{D})\phi(\theta_j; 0, \sigma_{\text{add}}^2)}, 1 \right).$$

**else**

        Choose one of the nonzero components (suppose it is the  $j$ th component) of  $\theta^{(i)}$  with probability  $1/|\mathcal{M}^{(i)}|$ .

        Set the  $j$ th component in  $\theta^{\text{new}}$  to zero.

        Set  $\theta^{i+1} = \theta^{\text{new}}$  with probability

$$\alpha_2 = \min \left( \frac{q(\theta^{\text{new}}|\mathbf{D})\phi(\theta_j; 0, \sigma_{\text{add}}^2)}{q(\theta^{(i)}|\mathbf{D})}, 1 \right).$$

**end if**

**(Within-model Step)**

    Set  $\mathcal{M}^{(i+1)}$  to be the model of  $\theta^{(i+1)}$ .

    Draw  $\theta^{\text{new}} \sim N(\theta^{(i+1)}, c \cdot \Xi_{\mathcal{M}^{(i+1)}})$ .

    Set  $\theta^{(i+1)} = \theta^{\text{new}}$  with probability

$$\alpha_3 = \min \left( \frac{q(\theta^{\text{new}}|\mathbf{D})}{q(\theta^{(i+1)}|\mathbf{D})}, 1 \right).$$

**end for**

---

### 3.2 Example 1: Correlated Binary Responses

The conditional mean  $\mu_{ij}(\theta) = E[Y_{ij}|X_{ij}]$  of the longitudinal binary response  $Y_{ij}$  is given by

$$\log \frac{\mu_{ij}(\theta)}{1 - \mu_{ij}(\theta)} = X_{ij}^\top \theta, \quad (10)$$

where  $i = 1, \dots, n$  and  $j = 1, \dots, s$ . In the following simulations, we fix the sample size  $n = 400$  and the cluster size  $s = 10$ . For  $X_{ij} = (X_{ij1}, \dots, X_{ijp})^\top$ , we consider two situations with  $p = 50$  and  $p = 100$ .  $X_{ij1}, \dots, X_{ijp}$  are generated independently from a uniform distribution on  $[-1, 1]$ . We also consider two sets of true parameter values,

$$\theta_0 = (1.5, -1.5, 1, -1, 0.5, -0.5, 0, 0, \dots, 0)$$

$$\theta_0 = (1.5, -1.5, 1.5, -1.5, 1, -1, 1, -1, 0.5, -0.5, 0.5, -0.5, 0, 0, \dots, 0)$$

with the number of nonzero components  $k_0 = 6$  and  $k_0 = 12$  respectively. Note that  $\theta_0$  contains weak signals 0.5 and  $-0.5$  and more nonzero components in the second setting. Similar to Wang et al. (2012) and Cho and Qu (2013), we use the R package `mvtBinaryEP` to generate the correlated binary responses with an exchangeable correlation structure with correlation coefficient  $\rho = 0.3$  within each cluster.

We compare the BGMM method to the frequentist penalized GEE method (PGEE) proposed by Wang et al. (2012). The PGEE solves a similar estimating equation to (4)

$$n^{-1} \sum_{i=1}^n \frac{\partial \mu_i(\theta)}{\partial \theta} \mathbf{S}_i^{-1} (Y_i - \mu_i(\theta)) - P_{\lambda_n}(\theta) = 0,$$

with an additional SCAD penalty  $P_\lambda(\theta) = (P_\lambda(\theta_1), \dots, P_\lambda(\theta_p))^\top$  and for  $k = 1, 2, \dots, p$ ,

$$P_\lambda(\theta_k) = \lambda_n \left[ 1(\theta_k \leq \lambda_n) + 1(\lambda_n < \theta_k \leq a\lambda_n) \frac{a\lambda_n - \theta_k}{(a-1)\lambda_n} \right].$$

The PGEE can be solved by an iterative Newton-Raphson algorithm as described in Wang et al. (2012). In our simulations, we fix  $a = 3.7$  and truncate the estimated coefficients to zero if  $|\hat{\theta}_k| \leq 10^{-3}$  ( $k = 1, 2, \dots, p$ ).  $\lambda_n$  is selected from the grid set  $\{0.01, 0.02, \dots, 0.2\}$  by 5-fold cross validation. We use an estimated correlation matrix for  $\mathbf{R}$  based on the sample, instead of varying the correlation structures in Wang et al. (2012). In fact, the sample estimation of  $\mathbf{R}$  is quite precise for the true  $\mathbf{R}$  in our  $p < n$  case.

For the BGMM method, the prior on  $\theta$  given a model  $\mathcal{M}$  is the product of independent normal densities

$$\pi(\theta|\mathcal{M}) = \prod_{j \in \mathcal{M}} \frac{1}{\sqrt{2\pi}\sigma_\theta} e^{-\frac{\theta_j^2}{2\sigma_\theta^2}}, \quad (11)$$

where we choose  $\sigma_\theta = 10$  for a large prior spread. It is straightforward to check that (11) satisfies Assumption 7.

For the prior on the model  $\mathcal{M}$ , we consider two different choices:

Prior 1:

$$\pi(\mathcal{M}) = \frac{1}{2^p - 1}; \quad (12)$$

Prior 2:

$$\pi(\mathcal{M}) = \frac{1}{p} \binom{p}{|\mathcal{M}|}^{-1}. \quad (13)$$

Prior 1 is the uniform prior over all the model spaces except the null space which we don't consider here (since there is no intercept term in the model). Prior 2 is similar to the prior used in Scott and Berger (2010) and Johnson and Rossell (2012), which first assigns uniform prior  $1/p$  on the model size, and then assigns a uniform prior over all models with the same size. Both priors satisfy our Assumption 8. Prior 2 can also be viewed as a hierarchical prior where each component of  $\theta$  enters the model independently with the same probability  $\nu$ , and then a hyper prior of Uniform[0,1] is imposed on  $\nu$ . In high dimensional settings when  $p$  gets large, the uniform prior (12) tends to favor models with size about  $p/2$  (see for example Chen and Chen 2008), which is undesirable for recovering the sparsity of  $\theta_0$ . Instead as shown in Scott and Berger (2010), the prior (13) favors the sparser models over the complicated models, by assigning geometrically decreasing prior probability as the model size increases from 0 to  $p/2$ . Therefore it induces a further penalization on the model size besides the incorporated BIC-type penalization in BGMM. From a frequentist point of view, Prior 1 and Prior 2 correspond to the two extreme cases of the extended BIC criterion in Chen and Chen (2008). In the following, we denote the BGMM method associated with them as BGMM1 and BGMM2, respectively.

In the algorithm of the BGMM method, the variance of proposal normal density is fixed at  $\sigma_{\text{add}} = 0.2$ . Our experiments with other values of  $\sigma_{\text{add}}$  (such as 0.05, 0.1, 0.15, 0.25) show that  $\sigma_{\text{add}} = 0.2$  is sufficient for exploring the full posterior of  $\theta$ , and the MCMC results such as the MAP model and posterior estimations are not sensitive to different

values of  $\sigma_{\text{add}}$ . For each of the 100 simulated datasets, we run one single chain with length  $3 \times 10^4$ , drop the first  $10^4$  iterations as burnin, and keep  $N = 1000$  MCMC samples from the remaining  $2 \times 10^4$  for every 20 iteration. All posterior summary statistics are calculated based this MCMC sample of size  $N = 1000$ .

As a benchmark, the PGEE method and the BGMM method are compared together with the naive method and the oracle method. The naive method estimates  $\theta$  by usual GEE without doing model selection, while for the oracle method, the true model is pretendedly known and  $\theta$  is estimated only on the nonzero components. We apply each method to the same dataset and repeat this process for 100 Monte Carlo replications. We compare three aspects of these methods: the model selection, the parameter estimation, and the prediction.

To evaluate the model selection performance, we consider the model selected by PGEE and the MAP model from BGMM, and report the proportion of times the method exact selecting (EX), underselecting (UN) and overselecting (OV) the nonzero components of  $\theta_0$ . We also report the true positives (TP, the average number of correctly selected nonzero components in  $\theta_0$ ), and the false positives (FP, the average number of selected nonzero components that are actually zero in  $\theta_0$ ).

For the estimation accuracy, we report the estimated mean square error (MSE)  $\sum_{m=1}^{100} \|\hat{\theta}_m - \theta_0\|^2 / (100k_0)$ , where  $\hat{\theta}_m$  is the  $\mathcal{M}$ th estimated parameter vector from the naive method, the oracle method, the PGEE method, and the posterior mean of  $\theta$  from the BGMM method.

For the prediction accuracy, we calculate the average MSE for the conditional mean  $\mu_{ij}$  (denoted by pMSE), defined as  $\sum_{i=1}^n \sum_{j=1}^s [\mu_{ij}(\hat{\theta}) - \mu_{ij}(\theta_0)]^2 / (ns)$ , for the naive method, the oracle method, and the PGEE method. The BGMM method uses the version averaged over the posterior sample  $\sum_{i=1}^n \sum_{j=1}^s \sum_{k=1}^N [\mu_{ij}(\theta_k) - \mu_{ij}(\theta_0)]^2 / (Nns)$ , where  $\theta_1, \dots, \theta_N$  are the MCMC sample of  $\theta$ .

As Table 1 indicates, both the frequentist PGEE method and our BGMM method have always successfully identified the nonzero components of  $\theta_0$  with no underselection. However, the PGEE performs much more conservative and has a serious overselection

problem in all the simulations settings, which is consistent with the findings in Cho and Qu (2013). It selects the true model for less than 30% of all time, and meanwhile over-selects about 4  $\sim$  8 extra redundant variables on average. In contrast, the BGMM MAP models with both the prior (12) and (13) are highly accurate when  $p = 50$ , and selects the true model for about 90% of all time with a much smaller false positives. When  $p = 100$ , the performance of BGMM1 deteriorates, but BGMM2 still maintains a high accuracy with over 90% of exact model selection. This is well explained by the regularization levels of the two priors (12) and (13), since the latter induces more penalization on the model size and fits the data better as  $p$  gets larger. For the estimation and prediction, it is clear that the naive GEE estimator with no model selection performs poorly in MSE and pMSE compared to the oracle estimator. Figure 1, 2 and 3 show that the MSE and pMSE for the BGMM method are comparable to those from the PGEE method, and tend to have smaller variation across difference simulations. The averaged levels of three MSEs from both BGMM methods are also slightly smaller than those from the PGEE estimator in most of the cases (Table 1), and they are all close to the MSE and pMSE from the oracle estimator. Overall, BGMM2 seems to be the best of all these methods besides the oracle. This has partly supported our theoretical results about the oracle properties of the BGMM method, in the sense that the posterior variance of BGMM is asymptotically the same as the variance of the oracle GMM estimator.

We also report the posterior probabilities of the MAP model  $\hat{\mathcal{M}}$  and the true model  $\mathcal{M}_0$  for BGMM1 and BGMM2 (Table 2), averaged over 100 simulated datasets. When  $p = 50$ , both methods yield reasonably large posterior probabilities for  $\hat{\mathcal{M}}$  and  $\mathcal{M}_0$ . For each method, the posterior of  $\hat{\mathcal{M}}$  and  $\mathcal{M}_0$  are close to each other. This is not surprising because of the high model selection accuracy for the BGMM2 MAP model. However when  $p = 100$ , the posterior probabilities of  $\hat{\mathcal{M}}$  and  $\mathcal{M}_0$  have dropped significantly for BGMM1, while the BGMM2 still performs well. This indicates that BGMM2 with the prior (13) can recover the sparsity better than BGMM1 with the prior (12) when  $p$  becomes large.



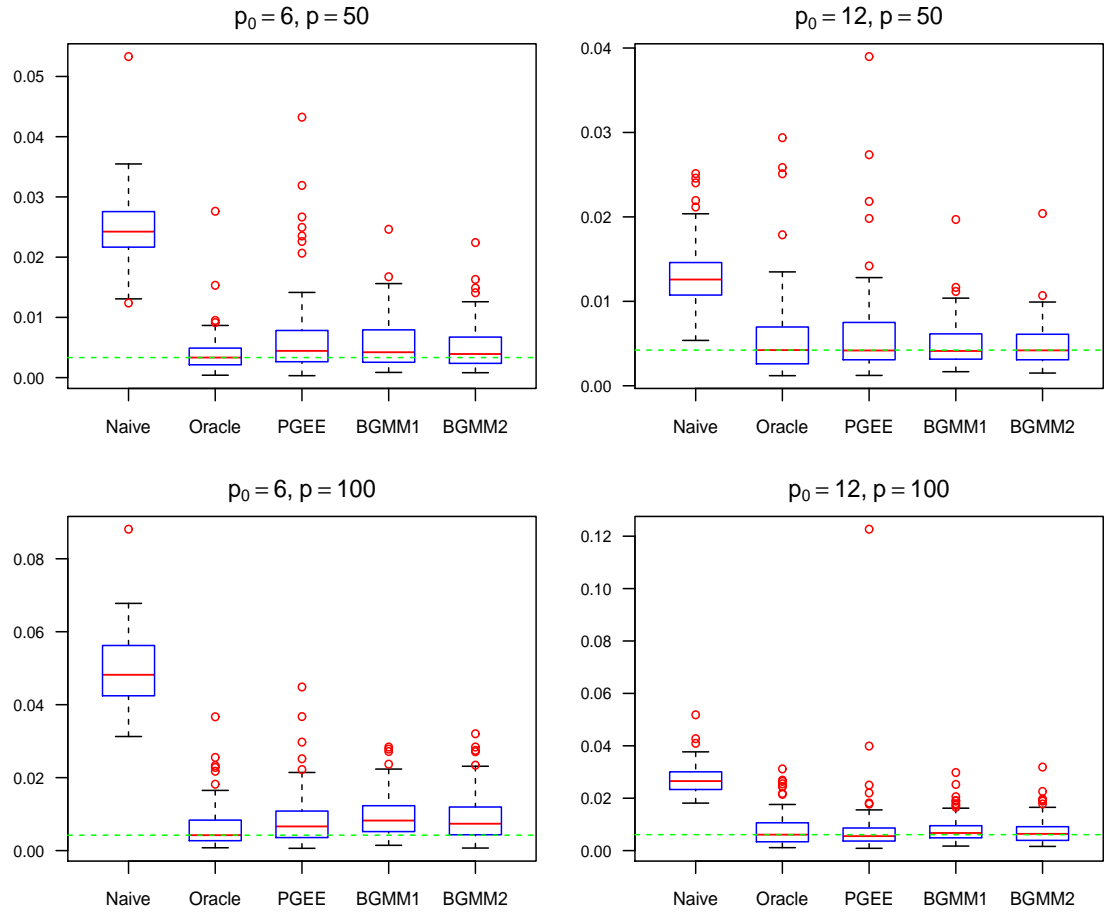


Figure 1: Boxplots for the MSE of  $\theta$  over 100 simulated datasets.

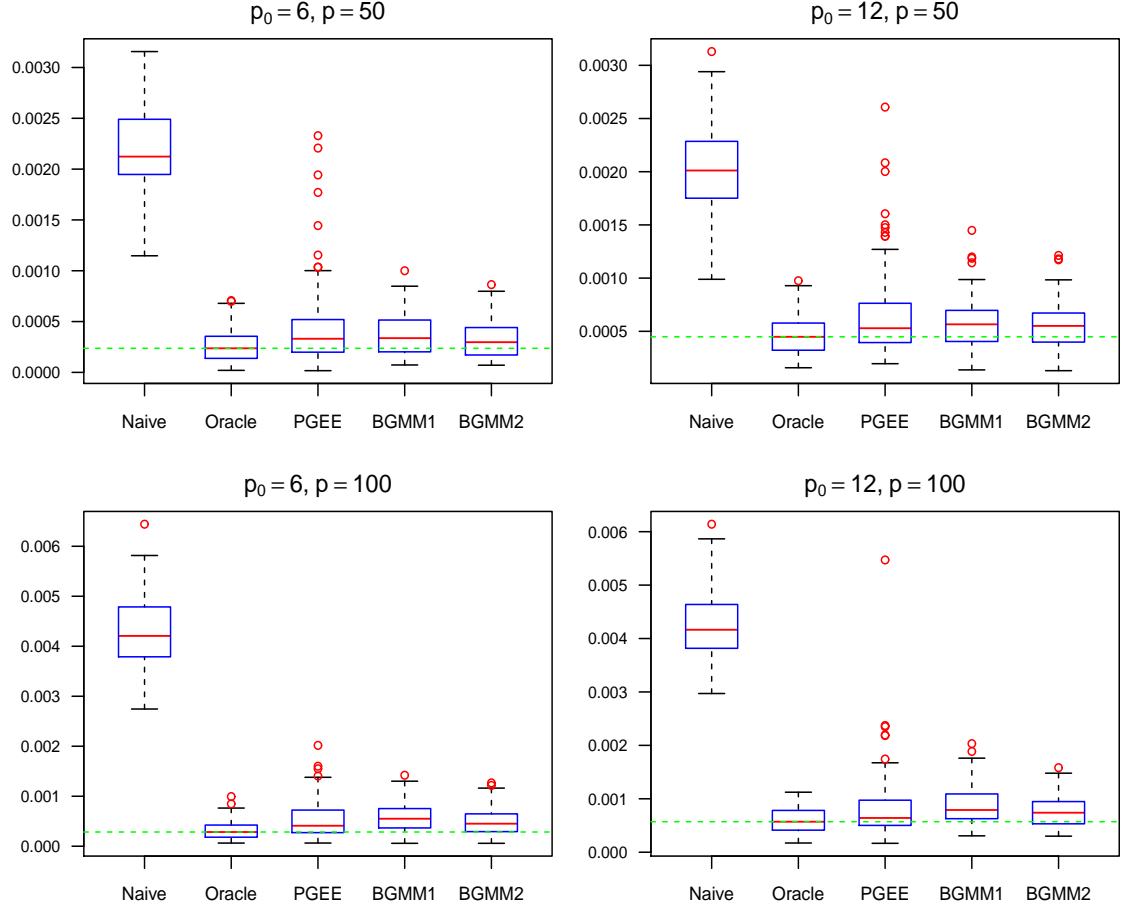


Figure 2: Boxplots for the MSE of  $\mu_{ij}(\theta)$  over 100 simulated datasets.

Table 1: Comparison of BGMM with PGEE for Binary Correlated Responses.  $k_0$  is the number of nonzero components in the true parameter  $\theta_0$ .  $p$  is the dimension of  $\theta_0$ .  $n$  is the sample size. Standard errors are shown in the parentheses. EX: exact selection; UN: under selection; OV: over selection; TP: true positives; FP: False positives. MSE: mean square error of  $\theta$ ; pMSE: prediction mean square error of  $\mu_{ij}(\theta)$ .

$k_0 = 6, p = 50, n = 400$							
	EX	UN	OV	TP	FP	MSE ( $\times 10^{-3}$ )	pMSE ( $\times 10^{-4}$ )
Naive	0	0	1	6	44	24.81 (0.57)	21.85 (0.42)
Oracle	1	0	0	6	0	4.15 (0.35)	2.65 (0.15)
PGEE	0.26	0	0.74	6	4.70	6.46 (0.68)	4.50 (0.42)
BGMM1	0.84	0	0.16	6	0.17	5.47 (0.39)	3.73 (0.22)
BGMM2	0.97	0	0.03	6	0.03	5.08 (0.38)	3.33 (0.20)
$k_0 = 12, p = 50, n = 400$							
	EX	UN	OV	TP	FP	MSE ( $\times 10^{-3}$ )	pMSE ( $\times 10^{-4}$ )
Naive	0	0	1	12	38	13.09 (0.35)	20.30 (0.39)
Oracle	1	0	0	12	0	5.53 (0.49)	4.80 (0.19)
PGEE	0.30	0	0.70	12	4.38	6.07 (0.54)	6.66 (0.42)
BGMM1	0.88	0	0.12	12	0.13	5.05 (0.29)	5.86 (0.23)
BGMM2	0.95	0	0.05	12	0.05	4.90 (0.28)	5.59 (0.21)
$k_0 = 6, p = 100, n = 400$							
	EX	UN	OV	TP	FP	MSE ( $\times 10^{-3}$ )	pMSE ( $\times 10^{-4}$ )
Naive	0	0	1	6	94	49.56 (0.90)	42.84 (0.67)
Oracle	1	0	0	6	0	6.46 (0.60)	3.18 (0.18)
PGEE	0.22	0	0.78	6	7.83	8.62 (0.73)	5.48 (0.40)
BGMM1	0.66	0	0.34	6	0.41	9.67 (0.60)	5.74 (0.28)
BGMM2	0.93	0	0.07	6	0.07	9.13 (0.66)	4.78 (0.25)
$k_0 = 12, p = 100, n = 400$							
	EX	UN	OV	TP	FP	MSE ( $\times 10^{-3}$ )	pMSE ( $\times 10^{-4}$ )
Naive	0	0	1	12	88	27.39 (0.55)	42.61 (0.61)
Oracle	1	0	0	12	0	8.09 (0.64)	5.90 (0.22)
PGEE	0.18	0	0.82	12	8.52	8.23 (1.28)	8.33 (0.64)
BGMM1	0.60	0	0.40	12	0.54	8.15 (0.51)	8.71 (0.35)
BGMM2	0.91	0	0.09	12	0.12	7.66 (0.51)	7.60 (0.29)

Table 2: Posterior probabilities averaged over 100 simulated datasets.  $k_0$  is the number of nonzero components in the true parameter  $\theta_0$ .  $p$  is the dimension of  $\theta_0$ .  $n$  is the sample size.  $q(\hat{\mathcal{M}}|\mathbf{D})$  is the BGMM posterior probability of the MAP model  $\hat{\mathcal{M}}$ .  $q(\mathcal{M}_0|\mathbf{D})$  is the BGMM posterior probability of the true model  $\mathcal{M}_0$ . The standard errors are shown in the parentheses. )

$[k_0, p, n]$	BGMM1		BGMM2	
	$q(\hat{\mathcal{M}} \mathbf{D})$	$q(\mathcal{M}_0 \mathbf{D})$	$q(\hat{\mathcal{M}} \mathbf{D})$	$q(\mathcal{M}_0 \mathbf{D})$
[6, 50, 400]	0.4879 (0.0140)	0.4534 (0.0181)	0.8439 (0.0118)	0.8386 (0.0135)
[12, 50, 400]	0.5631 (0.0140)	0.5286 (0.0193)	0.7738 (0.0140)	0.7618 (0.0169)
[6, 100, 400]	0.2063 (0.0108)	0.1726 (0.0129)	0.8067 (0.0141)	0.7776 (0.0206)
[12, 100, 400]	0.1961 (0.0102)	0.1624 (0.0121)	0.6732 (0.0180)	0.6559 (0.0210)

### 3.3 Example 2: Clinical Trial Dataset

The goal of this subsection is to illustrate the BGMM method for model selection through the application on the respiratory dataset used in Stokes et al. (2000), and compare it with the frequentist penalization method of PGEE. The dataset comes from a clinical trial that compares two treatments for a respiratory illness. 111 patients were randomly assigned to active treatment or placebo. The respiratory status of each patient was recorded at 4 visits at two clinical centers. The response variable is the binary respiratory status, coded as 1 for good outcome and 0 for poor outcome. There are also 5 covariates: clinical **center** (coded as 0 and 1 for two centers), **treatment** (coded as 0 for treatment and 1 for placebo), **gender** (coded as 0 for male and 1 for female), **age** at the beginning of the study (continuous), and **baseline** respiratory status (coded as 0 for poor and 1 for good).

Stokes et al. (2000) has analyzed this dataset using the standard GEE using the logit link function (10) as in Subsection 4.2. The **intercept** term was also included in the estimation. They have identified that **treatment** and **baseline** as two significant covariates, while **center** is on the borderline. Wang and Qu (2009) has further investigated the model selection problem using BIQIF (Bayesian information criterion based on the quadratic inference function) and also a data-driven smooth test for model checking. Their test indicates that the variable **age** has a quadratic effect on the response. After

adding the variable `age`<sup>2</sup>, they selected the model with `intercept`, `treatment`, `baseline`, `age` and `age`<sup>2</sup> as the best model.

We first compare the model selection among the 7 variables using both PGEE and BGMM. Both `age` and `age`<sup>2</sup> are standardized to roughly between -2 and 2. For the PGEE, we do the same 5-fold cross validation as in Subsection 4.2. We also consider the same two prior distributions on the models as in (12) and (13), and denote them by BGMM1 and BGMM2 respectively. Every BGMM posterior chain has length  $5 \times 10^4$ , with the first  $2 \times 10^4$  iterations dropped as burnin, and a subsample of length  $N = 1000$  is extracted from the rest of the chain for the final analysis, similar to Subsection 4.2. We will mainly focus on two models, which differ by two variables `age` and `age`<sup>2</sup>:

Model 1: `treatment`, `age`, `age`<sup>2</sup>, `baseline`;

Model 2: `treatment`, `baseline`.

We list out the model selected by PGEE and BGMM methods in Table 3. For the BGMM method, we list the top 5 models with the largest posterior probabilities. While PGEE and BGMM1 select Model 1, BGMM2 selects the sparser Model 2, but also assigns nonnegligible posterior probability to Model 1. Table 4 reports the estimates of  $\theta$  and the standard errors. For the PGEE method, the standard errors are calculated using the sandwich formula in Wang et al. (2012). For the BGMM methods, the estimates are the posterior means averaged over different models, and the standard errors are the posterior standard deviations. It is clear that BGMM2 tends to shrink the coefficients of `age` and `age`<sup>2</sup> towards zero. This is different from what is suggested by PGEE and leads us to consider further analysis on the effect of `age` and `age`<sup>2</sup>.

Table 3: Models selected by PGEE, BGMM1 and BGMM2 when  $p = 7$ .

	Selected Covariates	Posterior Prob.
PGEE	<code>treatment, age, age<sup>2</sup>, baseline</code>	-
BGMM1	<code>treatment, age, age<sup>2</sup>, baseline</code>	0.656
	<code>treatment, gender, age, age<sup>2</sup>, baseline</code>	0.153
	<code>center, treatment, age, age<sup>2</sup>, baseline</code>	0.083
	<code>treatment, baseline</code>	0.057
	<code>intercept, treatment, age, age<sup>2</sup>, baseline</code>	0.016
BGMM2	<code>treatment, baseline</code>	0.347
	<code>treatment, age, age<sup>2</sup>, baseline</code>	0.308
	<code>treatment, gender, age, age<sup>2</sup>, baseline</code>	0.139
	<code>center, treatment, age, age<sup>2</sup>, baseline</code>	0.069
	<code>center, treatment, baseline</code>	0.031

Table 4: Estimates of  $\theta$  and standard error when  $p = 7$ .

	PGEE	BGMM1	BGMM2
<code>intercept</code>	-	-0.002 (0.033)	-0.012 (0.098)
<code>center</code>	-	0.060 (0.215)	0.074 (0.236)
<code>treatment</code>	-1.160 (0.270)	-1.387 (0.478)	-1.336 (0.495)
<code>gender</code>	-	0.172 (0.447)	0.197 (0.501)
<code>age</code>	-1.872 (0.622)	-2.160 (0.904)	-1.442 (1.327)
<code>age<sup>2</sup></code>	1.210 (0.421)	1.339 (0.575)	0.889 (0.823)
<code>baseline</code>	2.072 (0.318)	2.329 (0.472)	2.359 (0.481)

If we include all the 7 variables in a usual GEE estimation, we will find that the coefficients of `age` and `age2` are indeed significantly different from zero, with individual Wald test p-values 0.0005 and 0.0008. However, we check further how the response of respiratory status changes with the variable `age` after adjusting for the effects of `treatment` and `baseline`. We divide the data into 4 groups according to the binary values of `treatment` and `baseline`. For each of the group, we run the usual GEE only with covariates `age` and `age2`, and draw the marginal lowess plot of the response v.s. `age` with 95% confidence bands. The results are summarized in Table 5, Figure 3 and 4. In fact, except in the last

group with `treatment=1` and `baseline=1`, both the p-values and the marginal lowess plots indicate that the respiratory status does not demonstrate any clear trend with `age`. Therefore when we fit the whole dataset and do model selection, the BGMM method has exhibited through the posterior model probabilities that the covariates `age` and `age2` are on the borderline. They either enter the model together in BGMM or they are not included in the selected model.

Table 5: GEE coefficients of `age` and `age2` according to the grouped values of `treatment` and `baseline`. The Wald test p-values are shown in the parentheses.

	age	age <sup>2</sup>
treatment=0, baseline=0	-2.457 (pval=0.027)	1.549 (pval=0.024)
treatment=0, baseline=1	-0.656 (pval=0.572)	0.572 (pval=0.445)
treatment=1, baseline=0	-1.559 (pval=0.247)	0.905 (pval=0.331)
treatment=1, baseline=1	-2.290 (pval=0.006)	1.405 (pval=0.019)

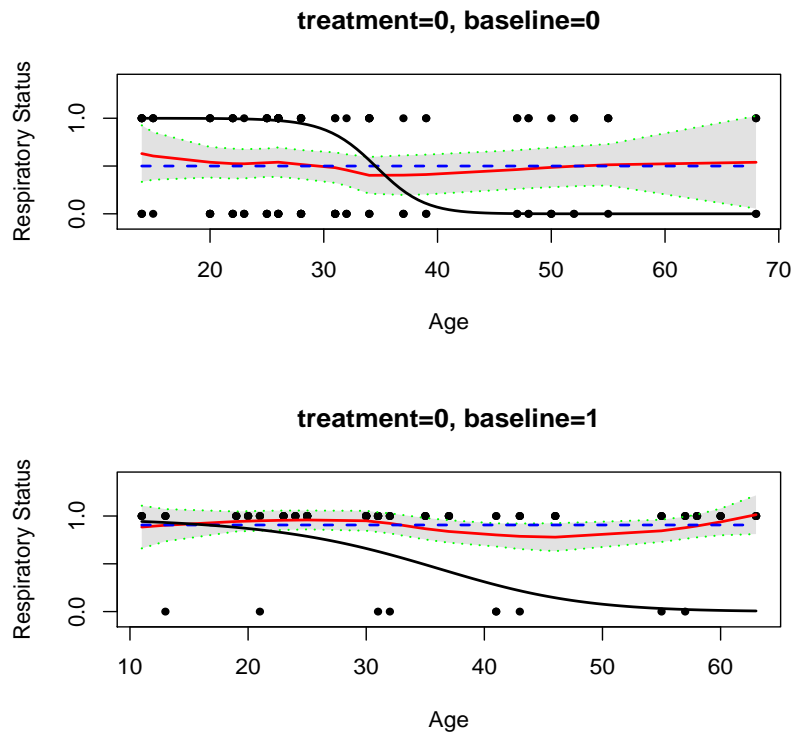


Figure 3: Marginal lowess plot with 95% confidence bands for the respiratory status v.s. `age`, with `treatment=0`. The blue horizontal dotted line is the sample mean. The solid black curve is the fitted logistic curve.

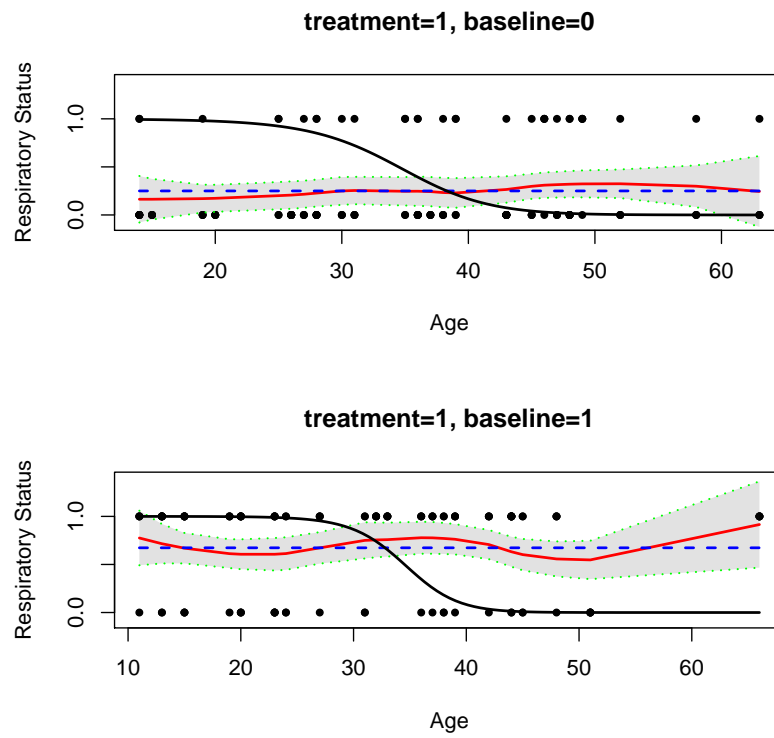


Figure 4: Marginal lowess plot with 95% confidence bands for the respiratory status v.s. age, with `treatment=1`. The blue horizontal dotted line is the sample mean. The solid black curve is the fitted logistic curve.



To compare the model selection performance of PGEE and BGMM under moderately high dimension, we artificially add 20 irrelevant covariates  $X_1, \dots, X_{20}$  ( $p = 27$ ) into the dataset and rerun the model selection procedure, which increases the dimension of  $\theta$  to  $p = 27$ . Each  $X_j$  is drawn independently from Uniform $[-1,1]$ . We give the results of one such realization in Table 6 and 7. PGEE and BGMM1 still select the same four covariates as in  $p = 7$  case, but the posterior probability for the BGMM1 MAP model has dropped compared to the  $p = 7$  case in Table 3. BGMM2 continues to select the sparser model with only **treatment** and **baseline**, and with a dominant posterior probability. The estimates of  $\theta$  are similar to Table 4, except that now BGMM2 dramatically shrinks the coefficients of **age** and **age**<sup>2</sup> to zero.

Table 6: Models selected by PGEE, BGMM1 and BGMM2 when  $p = 27$ .

	Selected Covariates	Posterior Prob.
PGEE	<b>treatment</b> , <b>age</b> , <b>age</b> <sup>2</sup> , <b>baseline</b>	-
BGMM1	<b>treatment</b> , <b>age</b> , <b>age</b> <sup>2</sup> , <b>baseline</b>	0.256
	<b>treatment</b> , <b>gender</b> , <b>age</b> , <b>age</b> <sup>2</sup> , <b>baseline</b>	0.087
	<b>treatment</b> , <b>age</b> , <b>age</b> <sup>2</sup> , <b>baseline</b> , $X_{25}$	0.084
	<b>center</b> , <b>treatment</b> , <b>age</b> , <b>age</b> <sup>2</sup> , <b>baseline</b> , $X_{16}$	0.045
	<b>treatment</b> , <b>baseline</b>	0.027
BGMM2	<b>treatment</b> , <b>baseline</b>	0.893
	<b>treatment</b> , <b>baseline</b> , $X_{26}$	0.010
	<b>treatment</b> , <b>age</b> , <b>baseline</b> , $X_{26}$	0.010
	<b>treatment</b> , <b>baseline</b> , $X_{12}$	0.009
	<b>treatment</b> , <b>baseline</b> , $X_{25}$	0.008

Table 7: Estimates of  $\theta$  and standard error when  $p = 27$ .

	PGEE	BGMM1	BGMM2
<b>intercept</b>	-	-0.002 (0.043)	-0.000 (0.002)
<b>center</b>	-	0.071 (0.263)	0.003 (0.048)
<b>treatment</b>	-1.150 (0.261)	-1.801 (0.773)	-1.283 (0.521)
<b>gender</b>	-	0.314 (0.732)	-0.000 (0.020)
<b>age</b>	-1.840 (0.602)	-2.376 (1.164)	-0.004 (0.039)
<b>age<sup>2</sup></b>	1.190 (0.408)	1.456 (0.746)	0.000 (0.006)
<b>baseline</b>	2.055 (0.308)	2.765 (0.807)	2.392 (0.595)

We further compare the performance of PGEE and BGMM after adding 0, 10 and 20 artificial irrelevant Uniform[-1,1] covariates ( $p = 7, 17, 27$ ), and repeat the process over 200 simulated datasets. Table 8 reports the number of true positives and the number of false positives for PGEE and BGMM averaged over 200 simulated datasets. Here the number of true positives (TP) refers to the number of selected covariates among the original 7 covariates. The number of false positives (FP) refers to the number of selected covariates among the artificially added covariates. It is clear that PGEE always has much larger TP and FP than the two BGMM methods, exhibiting serious overselection problem. In fact, of all 200 simulated datasets, PGEE only selects Model 1 for about half the time, while in the other cases it conservatively selects more irrelevant covariates. Also from Table 8, BGMM2 tends to be more parsimonious than BGMM1, as it selects **age** and **age<sup>2</sup>** less frequently with smaller marginal inclusion probabilities (Table 10). Table 9 reports the posterior probabilities of Model 1 and Model 2 for BGMM1 and BGMM2. When  $p$  is small (such as  $p = 7$ ), Model 1 and Model 2 have comparable posterior under both priors. But as  $p$  increases, BGMM2 favors Model 2 more, while the posteriors of both models have dropped significant for BGMM1. Moreover in Table 10, BGMM1 tends to include the seemingly redundant variable **gender** with an increasingly large probability. This implies that in this example, the regularization in BGMM1 may not be able to produce a MAP model with dominant posterior probability as  $p$  grows, while BGMM2 always succeeds in identifying the sparse Model 2.

Table 8: The true positives and false positives for PGEE and BGMM MAP model averaged over 200 simulated datasets..

		$p = 7$	$p = 17$	$p = 27$
TP	PGEE	5.31	5.00	4.54
	BGMM1	3.51	3.13	3.39
	BGMM2	2.99	2.42	2.32
FP	PGEE	0	1.44	1.49
	BGMM1	0	0.17	0.38
	BGMM2	0	0.10	0.18

Table 9: BGMM posterior probabilities of Model 1 and Model 2 averaged over 200 simulated datasets.

	$p = 7$	$p = 17$	$p = 27$
BGMM1	0.481/0.255	0.194/0.173	0.198/0.172
BGMM2	0.334/0.350	0.122/0.523	0.119/0.530

Table 10: Marginal BGMM posterior inclusion probabilities for the 7 original covariates, averaged over 200 simulated datasets.

		$p = 7$	$p = 17$	$p = 27$
<b>intercept</b>	BGMM1	0.027	0.034	0.054
	BGMM2	0.048	0.009	0.009
<b>center</b>	BGMM1	0.102	0.119	0.124
	BGMM2	0.128	0.038	0.026
<b>treatment</b>	BGMM1	0.998	0.985	0.973
	BGMM2	0.996	0.912	0.835
<b>gender</b>	BGMM1	0.141	0.196	0.265
	BGMM2	0.190	0.054	0.053
<b>age</b>	BGMM1	0.689	0.614	0.636
	BGMM2	0.602	0.247	0.232
<b>age<sup>2</sup></b>	BGMM1	0.680	0.596	0.609
	BGMM2	0.595	0.238	0.218
<b>baseline</b>	BGMM1	1.000	1.000	1.000
	BGMM2	1.000	1.000	1.000

## 4 Discussions

In this paper, we have studied some theoretical properties and applications of a Bayesian moment based model selection method. As we have commented, this method combines advantages of a Bayesian approach, such as the expressiveness of the posterior distribution and convenient MCMC algorithms for computation, with the model robustness of the moment based methods. We have formulated and proved the Bayesian oracle property of the proposed model selection method, which guarantees efficient posterior inference as if we knew which variables are truly relevant. We have studied the meaning of the quasi-posterior probabilities used in BGMM, which can be interpreted as the leading order large sample approximation to the true posterior probabilities conditional on the observed GMM estimator. The numerical performance of BGMM has been demonstrated by both simulated and real numerical examples.

Throughout the paper, we have assumed that a true model is a proper submodel of a full model, and we have considered the model selection among all the coordinate subspaces of  $\mathbb{R}^p$ . A different line of research by Hong and Preston (2012), has focused on the situation where two or more models are asymptotically equivalent in fit and all of them are misspecified. They have shown that in such cases, the model selection based on the GMM criterion function is not consistent if the models with the similar best fit are nonnested. We emphasize that such a situation in Hong and Preston (2012) does not apply to our model setup, because we have assumed the existence of a unique true model  $\mathcal{M}_0$  that is correctly specified while all the other models are not (Assumption 1 and 5(i)), and we have also considered an exhaustive model space with  $2^p$  candidate models. Therefore we arrive at the different conclusion that BGMM is consistent for model selection. Our assumption of a unique true model implies that the true parameter  $\theta_0$  contains some nonzero components and others are exactly zero, which may seem strong for practical purposes. In reality, the situation could be that all components of  $\theta_0$  are nonzero, with some larger in magnitude and many others much smaller (Jiang 2007). An interesting future work would be to study the theoretical properties in this more realistic setup.

We have only considered quasi-posterior constructed from the GMM based quasi-likelihood function. Many other alternatives, such as EL, GEL, and ETEL, can be formulated under a similar Bayesian framework, with possible interpretations of the induced quasi-Bayesian posterior. See for example, Chernozhukov and Hong (2003), Lazar (2003), Schennach (2005), etc. We conjecture that similar Bayesian asymptotic properties for model selection can be derived for these quasi-likelihoods.

Our theoretical results have been derived under increasing dimension with  $p \prec n$ . It would be interesting to generalize the results to ultra-high dimension with  $p \gg n$ . However, such generalization is highly nontrivial. Usually it is unrealistic to assume the existence of  $p$  exact moment conditions when  $p \gg n$ . Many moment conditions could be weak or even invalid. On one hand, we are supposed to include as many valid moments as possible, because it will decrease the asymptotic variance of the GMM estimator. On the other hand, the singularity issue from  $p > n$  makes it impossible to use the optimal GMM estimation for  $\theta$  as the number of valid moments gets too large, implying that for the estimation purpose, we can only use a sparse set of the valid moments. Recently Caner et al. (2013) and Chen and Liao (2013) have studied the moment selection problem using

the frequentist penalization method but still with  $p \prec n$ . For  $p \gg n$  in the Bayesian framework, a considerably different way to construct the quasi-likelihood function may be needed to handle this situation properly.

## Appendix A. Proofs and Lemmas

This appendix contains the proofs of all the theorems and corollaries in the main content. Some useful lemmas are presented with proofs. In the following, we write w.p.1  $-\eta$  short for “with probability at least  $1 - \eta$ ”.

**Lemma A.1.** (Belloni and Chernozhukov 2009) Let  $W_n(\mathbf{D}, \theta) = \bar{g}(\mathbf{D}, \theta) - Eg(D, \theta) - [\bar{g}(\mathbf{D}, \theta_0) - Eg(D, \theta_0)]$ . Then under Assumptions 2 and 4, uniformly for all  $\theta \in \Theta$ ,

$$\|W_n(\mathbf{D}, \theta)\| = O_p\left(\sqrt{\frac{p \log n}{n}} \|\theta - \theta_0\|^\alpha + n^{-1} p^{3/2} \log n\right).$$

**Proof:** The proof can be found in (A.10) of Belloni and Chernozhukov (2009). The  $L_2$  norm of the deviation  $W_n(\mathbf{D}, \theta)$  can be controlled using an empirical process result when the growth rate of  $p$  is specified in Assumption 2 and the moment  $g(D, \theta)$  satisfies Assumption 4. The only adaptation here is that the condition on VC dimension in ZE.1 of their paper is now replaced by the condition on the uniform covering number in Assumption 4(ii). The conclusion still holds according to the proof of Lemma 16 in Belloni et al. (2011). ■

**Lemma A.2.** Let  $\mathbf{G}_{\mathcal{M}}$  be the same as defined in Theorem 1. Define

$$S_{\mathcal{M}}(\mathbf{D}) = \exp\left\{-\frac{n}{2} \bar{g}(\mathbf{D}, \theta_0)^\top [\mathbf{V}_n^{-1} - \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}} (\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})^{-1} \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1}] \bar{g}(\mathbf{D}, \theta_0)\right\},$$

$$\bar{\theta}_{\mathcal{M},1} = \theta_{0,1} - (\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})^{-1} \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta_0),$$

where  $\theta_{0,1} \in \mathbb{R}^{|\mathcal{M}|}$  be the subvector of  $\theta_0$  restricted to  $\Theta(\mathcal{M})$ . Then under Assumptions 1-8, uniformly for all spaces  $\mathcal{M} \supseteq \mathcal{M}_0$ , for any fixed constant  $C > 0$ ,

$$\int_{B_0(C\epsilon_n) \cap \Theta(\mathcal{M})} e^{-\frac{n}{2} \bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta)} \pi(\theta | \mathcal{M}) d\theta$$

$$= (1 + o_p(1)) S_{\mathcal{M}}(\mathbf{D}) \int_{B_0(C\epsilon_n) \cap \Theta(\mathcal{M})} e^{-\frac{n}{2} (\theta_1 - \bar{\theta}_{\mathcal{M},1})^\top \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}} (\theta_1 - \bar{\theta}_{\mathcal{M},1})} \pi(\theta | \mathcal{M}) d\theta_1.$$

**Proof:** First of all, the  $L_2$  norm of  $\bar{g}(\mathbf{D}, \theta_0)$  satisfies for any  $C > 0$ ,

$$P\left(\|\bar{g}(\mathbf{D}, \theta_0)\| \geq C \sqrt{\frac{p}{n}}\right) \leq \frac{nE\|\bar{g}(\mathbf{D}, \theta_0)\|^2}{C^2 p} = \frac{\text{tr}(\text{Var}(g(D, \theta_0)))}{C^2 p} \leq \frac{\bar{\lambda}(\mathbf{V})}{C^2}, \quad (\text{A.1})$$

which implies that  $\|\bar{g}(\mathbf{D}, \theta)\| = O_p(\sqrt{p/n})$  since the eigenvalues of  $\mathbf{V} = \text{Var}[g(D, \theta_0)]$  are bounded above according to Assumption 6.

Second, for  $\theta \in B_0(C\epsilon_n) \cap \Theta(\mathcal{M})$ , let  $r_{\mathcal{M}}(\mathbf{D}, \theta) = \bar{g}(\mathbf{D}, \theta) - \bar{g}(\mathbf{D}, \theta_0) - \mathbf{G}_{\mathcal{M}}(\theta_1 - \theta_{0,1})$ . Then using second order Taylor expansion of  $Eg(\mathbf{D}, \theta)$  at  $\theta_0$  for  $\theta \in B_0(C\epsilon_n) \cap \Theta(\mathcal{M})$  and with all zero components of  $\theta$  excluded, we have

$$r_{\mathcal{M}}(\mathbf{D}, \theta) = \frac{1}{2} \mathbf{H}_M(\tilde{\theta}_1)(\theta_1 - \theta_{0,1}, \theta_1 - \theta_{0,1}) + W_n(\mathbf{D}, \theta),$$

where  $\tilde{\theta}_1$  is between  $\theta_1$  and  $\theta_{0,1}$  and  $\tilde{\theta} = (\tilde{\theta}_1^\top, 0)^\top$ ,  $\mathbf{H}_M$  is the submatrix of the second order derivative matrix  $\mathbf{H}$  restricted to  $\Theta(\mathcal{M})$ . By Assumption 5(iii), we have that

$$\|\mathbf{H}_M(\tilde{\theta}_1)(\theta_1 - \theta_{0,1}, \theta_1 - \theta_{0,1})\| \leq \sup_{\|u\|=1, \|v\|=1} \|\mathbf{H}(\tilde{\theta})[u, v]\| \|\theta_1 - \theta_{0,1}\|^2 \leq O(\sqrt{p}\epsilon_n^2).$$

Therefore, using Lemma A.1, we obtain that the order of  $r_{\mathcal{M}}(\mathbf{D}, \theta)$  on  $\theta \in B_0(C\epsilon_n) \cap \Theta(\mathcal{M})$  uniformly for all  $\mathcal{M} \supseteq \mathcal{M}_0$  is,

$$\|r_{\mathcal{M}}(\mathbf{D}, \theta)\| \leq O_p(n^{-1}p^{3/2} + (p/n)^{\frac{\alpha+1}{2}} \sqrt{\log n} + n^{-1}p^{3/2} \log n) = o_p((pn)^{-1/2}),$$

where the last equality is from the growth rate in Assumption 2. Then using decomposition  $\bar{g}(\mathbf{D}, \theta) = \mathbf{G}_{\mathcal{M}}(\theta_1 - \theta_{0,1}) + \bar{g}(\mathbf{D}, \theta_0) + r_{\mathcal{M}}(\mathbf{D}, \theta)$  we have

$$\begin{aligned} & \frac{n}{2} \bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta) \\ &= \frac{n}{2} \left\{ (\theta_1 - \bar{\theta}_{\mathcal{M},1})^\top (\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}}) (\theta_1 - \bar{\theta}_{\mathcal{M},1}) \right\} \\ &+ \frac{n}{2} \left\{ \bar{g}(\mathbf{D}, \theta_0)^\top [\mathbf{V}_n^{-1} - \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}} (\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})^{-1} \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1}] \bar{g}(\mathbf{D}, \theta_0) \right\} \\ &+ \frac{n}{2} \left\{ r_{\mathcal{M}}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} r_{\mathcal{M}}(\mathbf{D}, \theta) + 2r_{\mathcal{M}}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}} (\theta_1 - \theta_{0,1}) + 2r_{\mathcal{M}}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta_0) \right\} \end{aligned} \tag{A.2}$$

where  $\bar{\theta}$  is defined in the lemma. By Assumptions 5(ii) and 6, the eigenvalues of  $\mathbf{V}_n$  and  $\mathbf{G}^\top \mathbf{G}$  are bounded above and below w.p.a.1, so are the eigenvalues of any  $\mathbf{G}_{\mathcal{M}}^\top \mathbf{G}_{\mathcal{M}}$  since  $\mathbf{G}_{\mathcal{M}}$  is a submatrix of  $\mathbf{G}$ . Therefore on  $B_0(C\epsilon_n) \cap \Theta(\mathcal{M})$ ,  $\|r_{\mathcal{M}}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} r_{\mathcal{M}}(\mathbf{D}, \theta)\| \leq \underline{\lambda}(\mathbf{V}_n)^{-1} \|r_{\mathcal{M}}(\mathbf{D}, \theta)\|^2 = o_p((pn)^{-1})$ ,  $\|2r_{\mathcal{M}}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}} (\theta_1 - \theta_{0,1})\| \leq 2\underline{\lambda}(\mathbf{V}_n)^{-1} \bar{\lambda}(\mathbf{G}_{\mathcal{M}}^\top \mathbf{G}_{\mathcal{M}}) \cdot \|r_{\mathcal{M}}(\mathbf{D}, \theta)\| \cdot \|\theta - \theta_0\| \leq o_p((pn)^{-1/2}\epsilon_n) = o_p(n^{-1})$ ,  $\|2r_{\mathcal{M}}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta_0)\| \leq 2\underline{\lambda}(\mathbf{V}_n)^{-1} \cdot \|r_{\mathcal{M}}(\mathbf{D}, \theta)\| \cdot \|\bar{g}(\mathbf{D}, \theta_0)\| = o_p((pn)^{-1/2}(p/n)^{1/2}) = o_p(n^{-1})$ . These together imply that the last term in (A.2) is of order  $o_p(1)$ , and this holds uniformly for all  $\mathcal{M} \supseteq \mathcal{M}_0$ . The conclusion then follows if we define  $S_{\mathcal{M}}(\mathbf{D})$  as in the lemma, which does not depend on  $\theta$  and can be moved outside the integral.  $\blacksquare$



**Lemma A.3.** *Under Assumptions 1-8, there exists a constant  $C_1 > 0$ , such that uniformly for all spaces  $\mathcal{M} \supseteq \mathcal{M}_0$ , for any fixed constant  $C \geq C_1$  and all sufficiently large  $n$ ,*

$$\begin{aligned} & \int_{B_0(C\epsilon_n) \cap \Theta(\mathcal{M})} e^{-\frac{n}{2}(\theta_1 - \bar{\theta}_{\mathcal{M},1})^\top \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}} (\theta_1 - \bar{\theta}_{\mathcal{M},1})} \pi(\theta|\mathcal{M}) d\theta_1 \\ &= (2\pi/n)^{|\mathcal{M}|/2} [\det(\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})]^{-1/2} (\pi(\theta_0|\mathcal{M}) + o_p(1)), \end{aligned} \quad (\text{A.3})$$

where  $\bar{\theta}_{\mathcal{M},1}$  is defined in Lemma A.2.

**Proof:** First we let  $\mathbf{P}_{\mathcal{M}} = \mathbf{V}_n^{-1/2} \mathbf{G}_{\mathcal{M}} (\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})^{-1} \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1/2}$ , where  $\mathbf{V}_n^{1/2}$  is the symmetric positive definite square root of  $\mathbf{V}_n$ . Then  $\mathbf{P}_{\mathcal{M}}$  is idempotent and has eigenvalues 0 and 1. The difference between  $\bar{\theta}_{\mathcal{M},1}$  and  $\theta_{0,1}$  can be controlled by

$$\begin{aligned} \|\bar{\theta}_{\mathcal{M},1} - \theta_{0,1}\|^2 &= \|(\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})^{-1} \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta_0)\|^2 \\ &\leq \underline{\lambda}(\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})^{-1} \cdot \bar{g}(\mathbf{D}, \theta_0)^\top \mathbf{V}_n^{-1/2} \mathbf{P}_{\mathcal{M}} \mathbf{V}_n^{-1/2} \bar{g}(\mathbf{D}, \theta_0) \\ &\leq \bar{\lambda}(\mathbf{V}_n) \underline{\lambda}(\mathbf{G}^\top \mathbf{G})^{-1} \cdot \underline{\lambda}(\mathbf{V}_n)^{-1} \|\bar{g}(\mathbf{D}, \theta_0)\|^2. \end{aligned}$$

Since  $\|\bar{g}(\mathbf{D}, \theta_0)\| = O_p(\sqrt{p/n})$ , we know that  $\|\bar{\theta}_{\mathcal{M},1} - \theta_{0,1}\|$  is also  $O_p(\sqrt{p/n})$  since all the eigenvalues here are bounded. So for any small  $\eta > 0$ , we can pick  $C'$  sufficiently large, such that  $\|\bar{\theta}_{\mathcal{M},1} - \theta_{0,1}\| \leq C' \sqrt{p/n}$  w.p.1 -  $\eta$  and uniformly for all  $\mathcal{M} \supseteq \mathcal{M}_0$ .

Next we evaluate the integral on the left hand side of (A.3) on  $B_{\mathcal{M}}(C\epsilon_n) := \{\theta = (\theta_1^\top, 0)^\top \in \Theta(\mathcal{M}) : \|\theta_1 - \bar{\theta}_{\mathcal{M},1}\| \leq C\epsilon_n\}$  for a fixed  $C > 0$ . We observe that the integral takes the same form as a Gaussian random vector centered at  $\bar{\theta}_{\mathcal{M},1}$ . Define  $U \sim N(0, \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})$ . Then we have that there exists a large  $C''$ , such that when  $C \geq C''$ , w.p.1 -  $2\eta$  and uniformly for all  $\mathcal{M} \supseteq \mathcal{M}_0$ ,

$$\begin{aligned} & \int_{B_{\mathcal{M}}(C\epsilon_n)} e^{-\frac{n}{2}(\theta_1 - \bar{\theta}_{\mathcal{M},1})^\top \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}} (\theta_1 - \bar{\theta}_{\mathcal{M},1})} \pi(\theta|\mathcal{M}) d\theta_1 \\ &= (2\pi/n)^{|\mathcal{M}|/2} [\det(\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})]^{-1/2} P(\|U\| \leq C\sqrt{p}) \cdot (\pi(\theta_0|\mathcal{M}) + o(1)) \\ &= (2\pi/n)^{|\mathcal{M}|/2} [\det(\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})]^{-1/2} (\pi(\theta_0|\mathcal{M}) + o(1)), \end{aligned} \quad (\text{A.4})$$

where the  $o(1)$  depends on  $\eta$  (and hence on  $C$  and  $n$ ). In the first equality above, we have used Assumption 7(ii) that  $|\pi(\theta|\mathcal{M}) - \pi(\theta_0|\mathcal{M})| = o(1)$  uniformly over all  $\mathcal{M} \supseteq \mathcal{M}_0$  and all  $\theta \in B_0((C + C')\epsilon_n)$ , since w.p.1 -  $\eta$ ,  $B_{\mathcal{M}}(C\epsilon_n) \subseteq B_0((C + C')\epsilon_n)$ . In the second equality, we used the fact that the eigenvalues of  $\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}}$  are bounded in probability and by Chebyshev's inequality,  $\|U\| = O_p(\sqrt{p})$ . Hence we can pick a large  $C''$  such that for  $C \geq C''$ , w.p.1 -  $\eta$ ,  $P(\|U\| \leq C\sqrt{p}) = 1 + o(1)$ .

Finally we set  $C_1 = C' + C''$ . Then for  $C \geq C_1$ ,  $B_{\mathcal{M}}((C - C')\epsilon_n) \subseteq B_0(C\epsilon_n) \subseteq B_{\mathcal{M}}((C + C')\epsilon_n)$ , and  $C - C' \geq C''$  guarantees that (A.4) is satisfied w.p.  $1 - 2\eta$  and uniformly for all  $\mathcal{M} \supseteq \mathcal{M}_0$ . Therefore (A.3) follows since the integral on the left hand side of (A.3) can be bounded between the integrals on  $B_{\mathcal{M}}((C - C')\epsilon_n)$  and  $B_{\mathcal{M}}((C + C')\epsilon_n)$ , and both integrals satisfy (A.4).  $\blacksquare$

**Lemma A.4.** *Under Assumptions 1-8, there exists a constant  $C_2 > 0$ , such that uniformly for all spaces  $\mathcal{M} \supseteq \mathcal{M}_0$ , for all large constant  $C \geq C_2$  and all sufficiently large  $n$ , w.p.a.1,*

$$\begin{aligned} & \int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} e^{-\frac{n}{2}\bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1}\bar{g}(\mathbf{D}, \theta)} \pi(\theta|\mathcal{M}) d\theta \\ & \leq c_\pi \left( \frac{4\pi\bar{\lambda}}{n\delta_1^2} \right)^{|\mathcal{M}|/2} \exp\left(-\frac{C^2\delta_1^2}{16\bar{\lambda}}p\right) + \exp\left(-\frac{n}{4}\bar{\lambda}^{-1}\delta_0^2\right), \end{aligned}$$

where  $c_\pi$  is from Assumption 7(i), and  $\delta_0, \delta_1$  are from Assumption 5(i).

**Proof:** The proof uses similar techniques to the proof of Lemma 8 in Belloni and Chernozhukov (2009). Here we first directly cite part of the results from Belloni and Chernozhukov (2009), since they remain valid under our Assumptions 2, 4, and 5.

1. For any small  $\eta > 0$ , there exists a large  $C' > 0$ , such that  $\|Eg(D, \theta)\| > 8\|\bar{g}(\mathbf{D}, \theta_0)\|$  uniformly on  $\Theta \setminus B_0(C'\epsilon_n)$  w.p.  $1 - \eta$ .  $C'$  depends on  $\delta_0, \delta_1$  in Assumption 5(i) and  $\bar{\lambda}$  in Assumption 6.

2.  $\|W_n(\mathbf{D}, \theta)\| = o_p(\|Eg(D, \theta)\|)$  uniformly on  $\Theta \setminus B_0(C'\epsilon_n)$ . So for  $n$  sufficiently large,  $\|W_n(\mathbf{D}, \theta)\| \leq \|Eg(D, \theta)\|/8$  for all  $\theta \in \Theta \setminus B_0(C'\epsilon_n)$  w.p.  $1 - \eta$ .

Note that the two results above hold uniformly for all  $\theta \in \Theta(\mathcal{M}) \setminus B_0(C'\epsilon_n)$  and for all  $\mathcal{M} \supseteq \mathcal{M}_0$ . Therefore, we have

$$\begin{aligned} \|\bar{g}(\mathbf{D}, \theta)\| &= \|Eg(D, \theta) + \bar{g}(\mathbf{D}, \theta_0) + W_n(\mathbf{D}, \theta)\| \\ &\geq \left| \|Eg(D, \theta)\| - \|\bar{g}(\mathbf{D}, \theta_0)\| - \|W_n(\mathbf{D}, \theta)\| \right| \\ &\geq \frac{3}{4}\|Eg(D, \theta)\| \end{aligned}$$

uniformly for all  $\theta \in \Theta(\mathcal{M}) \setminus B_0(C'\epsilon_n)$ , all  $\mathcal{M} \supseteq \mathcal{M}_0$ , and all sufficiently large  $n$  w.p.  $1 - 2\eta$ .

Therefore, for  $C > C'$ , by Assumptions 1-8, we have

$$\int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} e^{-\frac{n}{2}\bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1}\bar{g}(\mathbf{D}, \theta)} \pi(\theta|\mathcal{M}) d\theta$$

$$\begin{aligned}
&\leq \int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} \exp \left\{ -\frac{n}{2} \bar{\lambda} (\mathbf{V}_n)^{-1} \cdot \frac{9}{16} \|Eg(D, \theta)\|^2 \right\} \pi(\theta|\mathcal{M}) d\theta \\
&\leq c_\pi \int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} \exp \left\{ -\frac{n}{4} \bar{\lambda}^{-1} \delta_1^2 \|\theta - \theta_0\|^2 \right\} d\theta \\
&\quad + \int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} \exp \left\{ -\frac{n}{4} \bar{\lambda}^{-1} \delta_0^2 \right\} \pi(\theta|\mathcal{M}) d\theta \\
&\leq c_\pi \left( \frac{4\pi \bar{\lambda}}{n\delta_1^2} \right)^{|\mathcal{M}|/2} P(\|U\| \geq C\sqrt{p}) + \exp \left( -\frac{n}{4} \bar{\lambda}^{-1} \delta_0^2 \right) \\
&\leq c_\pi \left( \frac{4\pi \bar{\lambda}}{n\delta_1^2} \right)^{|\mathcal{M}|/2} \exp \left( -\frac{C^2 \delta_1^2}{16\bar{\lambda}} p \right) + \exp \left( -\frac{n}{4} \bar{\lambda}^{-1} \delta_0^2 \right) \tag{A.5}
\end{aligned}$$

where in the second inequality we used Assumption 5(i), 6 and 7(i) and required  $n$  to be sufficiently large, in the third inequality we let  $U \sim N(0, \frac{2\bar{\lambda}}{\delta_1^2} \mathbf{I}_{|\mathcal{M}|})$ , applied the Gaussian concentration inequality and required  $C > C_2 = \max(2\sqrt{2\bar{\lambda}}/\delta_1, C')$ . The whole inequality holds uniformly for all  $\mathcal{M} \supseteq \mathcal{M}_0$ , and all sufficiently large  $n$ , w.p.  $1 - 2\eta$ .  $\blacksquare$

**Lemma A.5.** *Suppose Assumptions 1-8 holds. Then w.p.a.1, uniformly for all  $\mathcal{M} \supseteq \mathcal{M}_0$ ,*

$$\text{BF}_q[\mathcal{M} : \mathcal{M}_0] \asymp \left( \frac{2\pi}{n} \right)^{-\frac{|\mathcal{M}| - |\mathcal{M}_0|}{2}} \frac{S_{\mathcal{M}}(\mathbf{D})}{S_{\mathcal{M}_0}(\mathbf{D})} \cdot \frac{[\det(\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})]^{-1/2} \pi(\theta_0|\mathcal{M})}{[\det(\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})]^{-1/2} \pi(\theta_0|\mathcal{M}_0)}, \tag{A.6}$$

where  $S_{\mathcal{M}}(\mathbf{D})$  is defined in Lemma A.2. Moreover,

$$0 < \log \frac{S_{\mathcal{M}}(\mathbf{D})}{S_{\mathcal{M}_0}(\mathbf{D})} \asymp (|\mathcal{M}| - |\mathcal{M}_0|). \tag{A.7}$$

**Proof:** We first establish an approximation of the integral on the true model space  $\Theta(\mathcal{M}_0)$ . From Lemma A.2, A.3 and A.4, we can pick a large constant  $C > \max(C_1, C_2)$  such that for all sufficiently large  $n$ ,

$$\begin{aligned}
&\int_{\Theta(\mathcal{M}_0)} e^{-\frac{n}{2} \bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta)} \pi(\theta|\mathcal{M}_0) d\theta \\
&= (1 + o_p(1)) S_{\mathcal{M}_0}(\mathbf{D}) (2\pi/n)^{|\mathcal{M}_0|/2} [\det(\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})]^{-1/2} \pi(\theta_0|\mathcal{M}_0) \tag{A.8}
\end{aligned}$$

This is because the density at  $\theta_0$  on  $\mathcal{M}_0$  is lower bounded by  $e^{-c_0 k_0}$  by Assumption 7(iii), and hence the upper bound in Lemma A.4 is of smaller order compared to the right hand side of (A.3), which implies that the integral on the space  $\Theta(\mathcal{M}_0)$  is mostly concentrated on the neighborhood  $B(C\epsilon_n)$  and the outside part is negligible.

For any  $\mathcal{M} \supseteq \mathcal{M}_0$  and  $\mathcal{M} \neq \mathcal{M}_0$ , we can decompose the Bayes factor in two parts:

$$\text{BF}_q[\mathcal{M} : \mathcal{M}_0]$$

$$\begin{aligned}
&= \frac{\int_{\Theta(\mathcal{M}) \cap B_0(C\epsilon_n)} e^{-\frac{n}{2}\bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta)} \pi(\theta|\mathcal{M}) d\theta}{\int_{\Theta(\mathcal{M}_0)} e^{-\frac{n}{2}\bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta)} \pi(\theta|\mathcal{M}_0) d\theta} \\
&\quad + \frac{\int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} e^{-\frac{n}{2}\bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta)} \pi(\theta|\mathcal{M}) d\theta}{\int_{\Theta(\mathcal{M}_0)} e^{-\frac{n}{2}\bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta)} \pi(\theta|\mathcal{M}_0) d\theta} \\
&:= I_1 + I_2
\end{aligned} \tag{A.9}$$

Based on Lemma A.2, Lemma A.3 and (A.8),  $I_1$  can be bounded by

$$I_1 = \frac{(1 + o_p(1)) S_{\mathcal{M}}(\mathbf{D}) (2\pi/n)^{|\mathcal{M}|/2} [\det(\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})]^{-1/2} (\pi(\theta_0|\mathcal{M}) + o_p(1))}{(1 + o_p(1)) S_{\mathcal{M}_0}(\mathbf{D}) (2\pi/n)^{|\mathcal{M}_0|/2} [\det(\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})]^{-1/2} \pi(\theta_0|\mathcal{M}_0)}, \tag{A.10}$$

where the  $o_p(1)$  holds uniformly for all  $M \supset M_0$ . Now we analyze the term  $S_{\mathcal{M}}(\mathbf{D})/S_{\mathcal{M}_0}(\mathbf{D})$  and prove (A.7). According to the definition of  $S_{\mathcal{M}}(\mathbf{D})$  in Lemma A.2,

$$\begin{aligned}
\frac{S_{\mathcal{M}}(\mathbf{D})}{S_{\mathcal{M}_0}(\mathbf{D})} &= \exp \left\{ \frac{n}{2} \bar{g}(\mathbf{D}, \theta_0)^\top [\mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}} (\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})^{-1} \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \right. \\
&\quad \left. - \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0} (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1} \mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1}] \bar{g}(\mathbf{D}, \theta_0) \right\} \\
&= \exp \left\{ \frac{n}{2} \bar{g}(\mathbf{D}, \theta_0)^\top \mathbf{V}_n^{-1/2} (\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{M}_0}) \mathbf{V}_n^{-1/2} \bar{g}(\mathbf{D}, \theta_0) \right\},
\end{aligned} \tag{A.11}$$

where  $\mathbf{P}_{\mathcal{M}}$  is the projection matrix defined at the beginning of the proof of Lemma A.3. Given  $M \supset M_0$ ,  $\mathbf{P}_{\mathcal{M}} - \mathbf{P}_{\mathcal{M}_0}$  is semi-positive definite and idempotent, with trace  $|\mathcal{M}| - |\mathcal{M}_0|$ . Since by CLT,  $\sqrt{n} \mathbf{V}_n^{-1/2} \bar{g}(\mathbf{D}, \theta_0)$  converges in distribution to  $N(0, I_p)$ , it follows that  $2 \log \frac{S_{\mathcal{M}}(\mathbf{D})}{S_{\mathcal{M}_0}(\mathbf{D})}$  is asymptotically a  $\chi_{|\mathcal{M}| - |\mathcal{M}_0|}^2$  random variable. Hence (A.7) is proved.

For  $I_2$ , Lemma A.4, (A.8) and Assumption 7(i,iii) together yield

$$I_2 \leq \frac{c_\pi \left( \frac{4\pi\bar{\lambda}}{n\delta_1^2} \right)^{|\mathcal{M}|/2} \exp \left( -\frac{C^2\delta_1^2}{16\bar{\lambda}} p \right) + \exp \left( -\frac{n}{4} \bar{\lambda}^{-1} \delta_0^2 \right)}{(1 + o_p(1)) S_{\mathcal{M}_0}(\mathbf{D}) (2\pi/n)^{|\mathcal{M}_0|/2} [\det(\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})]^{-1/2} \pi(\theta_0|\mathcal{M}_0)}.$$

Hence we take the ratio of  $I_2$  to  $I_1$  in (A.10), and have

$$\begin{aligned}
\frac{I_2}{I_1} &\leq \frac{c_\pi \left( \frac{4\pi\bar{\lambda}}{n\delta_1^2} \right)^{|\mathcal{M}|/2} \exp \left( -\frac{C^2\delta_1^2}{16\bar{\lambda}} p \right) + \exp \left( -\frac{n}{4} \bar{\lambda}^{-1} \delta_0^2 \right)}{(1 + o_p(1)) S_{\mathcal{M}}(\mathbf{D}) (2\pi/n)^{|\mathcal{M}|/2} [\det(\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})]^{-1/2} \pi(\theta_0|\mathcal{M})} \\
&\leq \frac{(1 + o_p(1)) c_\pi \left( \frac{2\bar{\lambda}}{\delta_1^2} \right)^{|\mathcal{M}|/2} \exp \left( -\frac{C^2\delta_1^2}{16\bar{\lambda}} p \right) \cdot \left( \frac{\bar{\lambda}(\mathbf{G}^\top \mathbf{G})}{\bar{\lambda}(\mathbf{V}_n)} \right)^{|\mathcal{M}|/2}}{\exp \left\{ -\frac{n}{2} \bar{g}(\mathbf{D}, \theta_0)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta_0) \right\} \cdot e^{-c_0|\mathcal{M}|}},
\end{aligned} \tag{A.12}$$

where in the second inequality, we applied Assumption 7(iii), and also a lower bound on  $S_{\mathcal{M}}(\mathbf{D})$  using its definition in Lemma A.2. In fact, by (A.1)

$$\frac{n}{2} \bar{g}(\mathbf{D}, \theta_0)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta_0) = O_p(p).$$

So one can see that in (A.12), w.p.a.1, we can pick  $C$  in the numerator sufficiently large, such that  $I_2/I_1$  is arbitrarily small, since the exponential index  $|\mathcal{M}|$  cannot exceed  $p$ . Therefore, in (A.9),  $\text{BF}_q[\mathcal{M} : \mathcal{M}_0] = (1 + o_p(1))I_1$ , and the conclusion of (A.6) follows from this and (A.10).  $\blacksquare$

**Lemma A.6.** *Suppose Assumptions 1-8 holds. Then w.p.a.1, there exists a large constant  $C_1 > 0$ , such that uniformly over all  $\mathcal{M}$  with  $\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset$ ,*

$$\text{BF}_q[\mathcal{M} : \mathcal{M}_0] \leq \exp \left\{ -C_1 n \min_{j \in \mathcal{M}_0} \theta_{0,j}^2 \right\}. \quad (\text{A.13})$$

**Proof:** First we observe that if  $\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset$ , i.e.  $\mathcal{M}$  misses at least one component of the true model  $\mathcal{M}_0$ , then for any  $\theta \in \Theta(\mathcal{M})$ , it must hold that  $\|\theta - \theta_0\| \geq \min_{j: \theta_{0,j} \neq 0} |\theta_{0,j}|$ . By Assumption 3, there exists a sequence  $t_n \rightarrow \infty$  such that  $\min_{j \in \mathcal{M}_0} |\theta_{0,j}| = \sqrt{\log n} t_n \epsilon_n$ . Therefore for  $n$  sufficiently large, the whole space  $\Theta(\mathcal{M})$  is outside the neighborhood  $B_0(\sqrt{\log n} t_n \epsilon_n)$ . Similar to the derivation of (A.5), we can bound the marginal probability  $q(\mathbf{D}|\mathcal{M})$  by

$$\begin{aligned} & \int_{\Theta(\mathcal{M})} e^{-\frac{n}{2} \bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta)} \pi(\theta|\mathcal{M}) d\theta \\ &= \int_{\Theta(\mathcal{M}) \setminus B_0(\sqrt{\log n} t_n \epsilon_n)} e^{-\frac{n}{2} \bar{g}(\mathbf{D}, \theta)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta)} \pi(\theta|\mathcal{M}) d\theta \\ &\leq c_\pi \left( \frac{4\pi \bar{\lambda}}{n \delta_1^2} \right)^{|\mathcal{M}|/2} \exp \left( -\frac{C^2 \delta_1^2}{16 \bar{\lambda}} t_n^2 p \log n \right) + \exp \left( -\frac{n}{4} \bar{\lambda}^{-1} \delta_0^2 \right) \end{aligned} \quad (\text{A.14})$$

for  $C$  sufficiently large w.p.a.1.

Therefore, by using the approximation (A.8), we have that w.p.a.1,

$$\begin{aligned} & \text{BF}_q[\mathcal{M} : \mathcal{M}_0] \\ &\leq \frac{c_\pi \left( \frac{4\pi \bar{\lambda}}{n \delta_1^2} \right)^{|\mathcal{M}|/2} \exp \left( -\frac{C^2 \delta_1^2}{16 \bar{\lambda}} t_n^2 p \log n \right) + \exp \left( -\frac{n}{4} \bar{\lambda}^{-1} \delta_0^2 \right)}{(1 + o_p(1)) S_{\mathcal{M}_0}(\mathbf{D}) (2\pi/n)^{|\mathcal{M}_0|/2} [\det(\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})]^{-1/2} \pi(\theta_0|\mathcal{M}_0)} \\ &\leq \frac{(1 + o_p(1)) c_\pi \left( \frac{2\bar{\lambda}}{\delta_1^2} \right)^{|\mathcal{M}|/2} \exp \left( -\frac{C^2 \delta_1^2}{16 \bar{\lambda}} t_n^2 p \log n \right)}{\exp \left\{ -\frac{n}{2} \bar{g}(\mathbf{D}, \theta_0)^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta_0) \right\} \left( \frac{\lambda(\mathbf{V}_n)}{\bar{\lambda}(\mathbf{G}^\top \mathbf{G})} \right)^{k_0/2} e^{-c_0 k_0}}, \end{aligned}$$

where in the last inequality, we did the same as in (A.12), and the second term on the numerator is absorbed into the first term because  $t_n^2 p \log n = n \min_{j \in \mathcal{M}_0} \theta_{0,j}^2 \leq n$  by Assumption 3. Now because the term  $t_n^2 p \log n$  dominates all the other terms in the exponential, the conclusion follows by choose appropriate  $C_1 > 0$ .  $\blacksquare$

**Proof of Theorem 1 (i):**

We combine the results from Lemma A.5 and Lemma A.6, and show that  $\sum_{\mathcal{M} \neq \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] \rightarrow 0$  as  $n \rightarrow \infty$  w.p.a.1, since this is equivalent to  $q(\mathcal{M}_0 | \mathbf{D}) \rightarrow 1$  w.p.a.1. To prove this, it suffices to show  $\sum_{\mathcal{M} : \mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] \rightarrow 0$  and  $\sum_{\mathcal{M} : \mathcal{M} \supseteq \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] \rightarrow 0$  respectively. For those models with  $\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset$ , using Lemma A.6 and Assumption 8(ii), we have

$$\begin{aligned} & \sum_{\mathcal{M} : \mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] \\ & \leq \sum_{\mathcal{M} : \mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset} \frac{\pi(\mathcal{M})}{\pi(\mathcal{M}_0)} \exp \left\{ -C_1 n \min_{j \in \mathcal{M}_0} \theta_{0,j}^2 \right\} \\ & \leq 2^p \cdot \exp \left\{ r_2 p \log n - C_1 n \min_{j \in \mathcal{M}_0} \theta_{0,j}^2 \right\} \rightarrow 0, \end{aligned}$$

since by Assumption 3,  $p \log n \prec n \min_{j \in \mathcal{M}_0} \theta_{0,j}^2$ .

For those models with  $\mathcal{M} \supseteq \mathcal{M}_0$ , using Lemma A.5, we can pick a large constant  $C_2 > 0$ , such that uniformly over all these models, for all sufficiently large  $n$ , w.p.a.1

$$\text{BF}_q[\mathcal{M} : \mathcal{M}_0] \leq (C_2 n)^{-\frac{|\mathcal{M}| - |\mathcal{M}_0|}{2}}.$$

Therefore, by Assumption 8(i), and  $p \prec n^{1/2}$ , we have

$$\begin{aligned} & \sum_{\mathcal{M} : \mathcal{M} \supseteq \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] \\ & \leq \sum_{\mathcal{M} : \mathcal{M} \supseteq \mathcal{M}_0} \frac{\pi(\mathcal{M})}{\pi(\mathcal{M}_0)} (C_2 n)^{-\frac{|\mathcal{M}| - |\mathcal{M}_0|}{2}} \\ & \leq \sum_{k=k_0+1}^p \binom{p-k_0}{k-k_0} r_1 (C_2 n)^{-\frac{k-k_0}{2}} \\ & = r_1 \left[ (1 + (C_2 n)^{-1/2})^{p-k_0} - 1 \right] \\ & \leq r_1 (e^{\frac{p-k_0}{\sqrt{C_2 n}}} - 1) \rightarrow 0. \end{aligned}$$

So the proof is complete. ■

**Proof of Theorem 1 (ii):**

In the conclusion of part (ii), the first integral can be rewritten as

$$\begin{aligned} & \int_A q(\theta | \mathbf{D}) d\theta \\ & = \frac{\sum_{\mathcal{M}} \pi(\mathcal{M}) \int_{A \cap \Theta(\mathcal{M})} q(\mathbf{D} | \theta, \mathcal{M}) \pi(\theta | \mathcal{M}) d\theta}{\sum_{\mathcal{M}} \pi(\mathcal{M}) \int_{\Theta(\mathcal{M})} q(\mathbf{D} | \theta, \mathcal{M}) \pi(\theta | \mathcal{M}) d\theta} \end{aligned}$$

$$= \frac{\sum_{\mathcal{M} \neq \mathcal{M}_0} \pi(\mathcal{M}) \int_{A \cap \Theta(\mathcal{M})} q(\mathbf{D}|\theta, \mathcal{M}) \pi(\theta|\mathcal{M}) d\theta + \pi(\mathcal{M}_0) \int_{A \cap \Theta(\mathcal{M}_0)} q(\mathbf{D}|\theta, \mathcal{M}) \pi(\theta|\mathcal{M}_0) d\theta}{\sum_{\mathcal{M} \neq \mathcal{M}_0} \pi(\mathcal{M}) \int_{\Theta(\mathcal{M})} q(\mathbf{D}|\theta, \mathcal{M}) \pi(\theta|\mathcal{M}) d\theta + \pi(\mathcal{M}_0) \int_{\Theta(\mathcal{M}_0)} q(\mathbf{D}|\theta, \mathcal{M}) \pi(\theta|\mathcal{M}_0) d\theta}.$$

Therefore if we divide the numerator and denominator by  $\pi(\mathcal{M}_0) \int_{\Theta(\mathcal{M}_0)} q(\mathbf{D}|\theta, \mathcal{M}) \pi(\theta|\mathcal{M}_0) d\theta$ , we have

$$\frac{\int_{A \cap \Theta(\mathcal{M}_0)} \tilde{q}(\theta_1|\mathbf{D}) d\theta_1}{\sum_{\mathcal{M} \neq \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] + 1} \leq \int_A q(\theta|\mathbf{D}) d\theta \leq \frac{\sum_{\mathcal{M} \neq \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] + \int_{A \cap \Theta(\mathcal{M}_0)} \tilde{q}(\theta_1|\mathbf{D}) d\theta_1}{\sum_{\mathcal{M} \neq \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] + 1},$$

where  $\tilde{q}(\theta_1|\mathbf{D}) = \frac{q(\mathbf{D}|\theta, \mathcal{M}) \pi(\theta|\mathcal{M}_0)}{\int_{\Theta(\mathcal{M}_0)} q(\mathbf{D}|\theta, \mathcal{M}) \pi(\theta|\mathcal{M}_0) d\theta}$  is the conditional quasi-posterior on the true model space  $\mathcal{M}_0$  and  $\theta_1$  represents the nonzero components of  $\theta$  in the model  $\mathcal{M}_0$ . According to the model selection consistency of Theorem 1 part (i),  $\sum_{\mathcal{M} \neq \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] \rightarrow 0$  w.p.a.1. Hence

$$\frac{\int_{A \cap \Theta(\mathcal{M}_0)} \tilde{q}(\theta_1|\mathbf{D}) d\theta_1}{1 + o_p(1)} \leq \int_A q(\theta|\mathbf{D}) d\theta \leq \frac{\int_{A \cap \Theta(\mathcal{M}_0)} \tilde{q}(\theta_1|\mathbf{D}) d\theta_1 + o_p(1)}{1 + o_p(1)}.$$

Since  $\int_{\Theta(\mathcal{M}_0)} \tilde{q}(\theta_1|\mathbf{D}) d\theta_1 = 1$  and  $\int_{\Theta} q(\theta|\mathbf{D}) d\theta = 1$ , and the  $o_p(1)$  does not depend on the set  $A$ , the inequality above implies that for all set  $A \subseteq \Theta$ , w.p.a.1.

$$\left| \int_A q(\theta|\mathbf{D}) d\theta - \int_{A \cap \Theta(\mathcal{M}_0)} \tilde{q}(\theta_1|\mathbf{D}) d\theta_1 \right| \rightarrow 0.$$

Therefore, to show part (ii) of Theorem 1, it suffices to show

$$\sup_{A \subseteq \Theta} \left| \int_{A \cap \Theta(\mathcal{M}_0)} \tilde{q}(\theta_1|\mathbf{D}) d\theta_1 - \int_{A \cap \Theta(\mathcal{M}_0)} \phi(\theta_1; \bar{\theta}_{\mathcal{M}_0, 1}, (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1}/n) d\theta_1 \right| \rightarrow 0, \text{ w.p.a.1.} \quad (\text{A.15})$$

Note that now the densities  $\tilde{q}$  and  $\phi$  are defined on the same support  $\Theta(\mathcal{M}_0)$ , so the rest is a standard proof of Bayesian CLT similar to Belloni and Chernozhukov (2009). Using the decomposition (A.2) in Lemma A.2, Lemma A.3 and Lemma A.4, we have

$$\begin{aligned} & \int_{A \cap \Theta(\mathcal{M}_0)} \tilde{q}(\theta_1|\mathbf{D}) d\theta_1 \\ &= \frac{S_{\mathcal{M}_0}(\mathbf{D}) \int_{A \cap \Theta(\mathcal{M}_0)} e^{-\frac{n}{2}(\theta_1 - \bar{\theta}_{\mathcal{M}_0, 1})^\top \mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0} (\theta_1 - \bar{\theta}_{\mathcal{M}_0, 1}) + o_p(1)} \pi(\theta_1|\mathcal{M}_0) d\theta_1}{(1 + o_p(1)) S_{\mathcal{M}_0}(\mathbf{D}) (2\pi/n)^{|\mathcal{M}_0|/2} [\det(\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})]^{-1/2} \pi(\theta_0|\mathcal{M}_0)} \\ &= \frac{(1 + o_p(1)) \int_{A \cap B_0(C\epsilon_n) \cap \Theta(\mathcal{M}_0)} e^{-\frac{n}{2}(\theta_1 - \bar{\theta}_{\mathcal{M}_0, 1})^\top \mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0} (\theta_1 - \bar{\theta}_{\mathcal{M}_0, 1})} \pi(\theta_1|\mathcal{M}_0) d\theta_1}{(2\pi/n)^{|\mathcal{M}_0|/2} [\det(\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})]^{-1/2} \pi(\theta_0|\mathcal{M}_0)} \\ &= (1 + o_p(1)) \int_{A \cap B_0(C\epsilon_n) \cap \Theta(\mathcal{M}_0)} \phi(\theta_1; \bar{\theta}_{\mathcal{M}_0, 1}, (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1}/n) d\theta_1 \\ &= \int_{A \cap B_0(C\epsilon_n) \cap \Theta(\mathcal{M}_0)} \phi(\theta_1; \bar{\theta}_{\mathcal{M}_0, 1}, (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1}/n) d\theta_1 + o_p(1) \end{aligned} \quad (\text{A.16})$$

The numerator of the second equality shrinks the range of integral to within the neighborhood  $B_0(C\epsilon_n)$  because the integral outside  $B_0(C\epsilon_n)$  is of  $o_p(1)$  compared to the denominator, according to the approximation in (A.8). In the third equality, we used Assumption 7(iii) and have that on  $B_0(C\epsilon_n) \cap \Theta(\mathcal{M}_0)$ ,  $\frac{\pi(\theta|\mathcal{M}_0)}{\pi(\theta_0|\mathcal{M}_0)} = 1 + o(1)$ . The  $o_p(1)$  in the last expression does not depend on the set  $A$ . Therefore (A.15) holds and this completes the proof.  $\blacksquare$

### Proof of Corollary 1:

Given  $\|\bar{\theta}_{\mathcal{M}_0,1} - \hat{\theta}_{\mathcal{M}_0,1}\| = O_p(p^{3/2}/n)$  in Assumption 9, it follows that for any  $\theta_1 \in B_0(C\epsilon_n) \cap \Theta(\mathcal{M}_0)$ ,

$$\begin{aligned} & \frac{n}{2}(\theta_1 - \bar{\theta}_{\mathcal{M}_0,1})^\top \mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0} (\theta_1 - \bar{\theta}_{\mathcal{M}_0,1}) \\ &= \frac{n}{2}(\theta_1 - \hat{\theta}_{\mathcal{M}_0,1})^\top \mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0} (\theta_1 - \hat{\theta}_{\mathcal{M}_0,1}) + n(\hat{\theta}_{\mathcal{M}_0,1} - \bar{\theta}_{\mathcal{M}_0,1})^\top \mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0} (\theta_1 - \hat{\theta}_{\mathcal{M}_0,1}) \\ & \quad + \frac{n}{2}(\hat{\theta}_{\mathcal{M}_0,1} - \bar{\theta}_{\mathcal{M}_0,1})^\top \mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0} (\hat{\theta}_{\mathcal{M}_0,1} - \bar{\theta}_{\mathcal{M}_0,1}) \\ &= \frac{n}{2}(\theta_1 - \hat{\theta}_{\mathcal{M}_0,1})^\top \mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0} (\theta_1 - \hat{\theta}_{\mathcal{M}_0,1}) + O_p(p^2/n^{1/2}) + O_p(p^3/n) \\ &= \frac{n}{2}(\theta_1 - \hat{\theta}_{\mathcal{M}_0,1})^\top \mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0} (\theta_1 - \hat{\theta}_{\mathcal{M}_0,1}) + o_p(1), \end{aligned}$$

where the last equality follows from the growth rate of  $p$ . Therefore in the display of (A.16), one can substitute each  $\theta_1 - \bar{\theta}_{\mathcal{M}_0,1}$  by  $\theta_1 - \hat{\theta}_{\mathcal{M}_0,1}$ , and get

$$\int_{A \cap \Theta(\mathcal{M}_0)} \tilde{q}(\theta_1 | \mathbf{D}) d\theta_1 = \int_{A \cap B_0(C\epsilon_n) \cap \Theta(\mathcal{M}_0)} \phi(\theta_1; \hat{\theta}_{\mathcal{M}_0,1}, (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1}/n) d\theta_1 + o_p(1),$$

where  $o_p(1)$  does not depend on  $A$ .  $\blacksquare$

### Proof of Theorem 2:

In the following, for a generic random variable  $D = (Y, \mathbf{X})^\top$  (independent of the sample  $\mathbf{D}$ ), we omit the subscript  $i$  in  $X_{ij}$  and write  $X_{\cdot j}$  to represent a generic  $p$ -dimensional covariate vector measured at time  $j$ , for  $j = 1, 2, \dots, s$ . Define  $\mu_{\cdot}(\theta) = (\mu(X_{\cdot 1}^\top \theta), \dots, \mu(X_{\cdot s}^\top \theta))^\top$ ,  $\mathbf{B}(\theta) = \frac{\partial \mu_{\cdot}(\theta)}{\partial \theta} = (\dot{\mu}(X_{\cdot 1}^\top \theta)X_{\cdot 1}, \dots, \dot{\mu}(X_{\cdot s}^\top \theta)X_{\cdot s})^\top$ , and  $\mathbf{S}(\theta) = \mathbf{A}(\theta)^{1/2} \mathbf{R} \mathbf{A}(\theta)^{1/2}$  (if  $X$  has sample index  $i$ , then we use the notation  $\mathbf{S}_i(\theta)$ ). So the generic moment condition can be written as  $g(D, \theta) = \mathbf{B}(\tilde{\theta})^\top \mathbf{S}(\tilde{\theta})^{-1} (Y - \mu_{\cdot}(\theta))$ , where  $\tilde{\theta}$  is a preliminary estimator that solves  $\sum_{i=1}^n \mathbf{X}_i (Y_i - \mu_i(\theta)) = 0$ . Similar to the proof of (3.3) in Wang (2011), one can show that given the conditions (1)-(3),  $\|\tilde{\theta} - \theta_0\| = O_p(\sqrt{p/n})$ . For simplicity we omit the proof of this relation here.

Suppose that the constant upper and lower bounds for  $\phi(X_{ij}^\top \theta_0)$  in Condition (3) are  $\bar{\phi}$  and  $\underline{\phi}$  respectively. Since w.p.a.1,  $\|\tilde{\theta} - \theta_0\| \leq C\epsilon_n$ ,  $\mathbf{S}_i(\tilde{\theta}) = \mathbf{A}_i(\tilde{\theta})^{1/2} \mathbf{R} \mathbf{A}_i(\tilde{\theta})^{1/2}$  and



$\mathbf{A}_i(\theta) = \text{diag} \{ \phi(X_{i1}^\top \theta), \dots, \phi(X_{is}^\top \theta) \}$ , we know that w.p.a.1, the eigenvalues of  $\mathbf{S}_i(\tilde{\theta})$  can be bounded as

$$\begin{aligned}\bar{\lambda}(\mathbf{S}_i(\tilde{\theta})) &\leq \bar{\lambda}(\mathbf{R}) \bar{\lambda}(\mathbf{A}_i(\tilde{\theta})) \leq \bar{\lambda}(\mathbf{R}) \text{tr}(\mathbf{A}_i(\tilde{\theta})) \leq s\bar{\phi} \bar{\lambda}(\mathbf{R}) \\ \underline{\lambda}(\mathbf{S}_i(\tilde{\theta})) &\geq \underline{\lambda}(\mathbf{R}) \underline{\lambda}(\mathbf{A}_i(\tilde{\theta})) \geq \underline{\phi} \underline{\lambda}(\mathbf{R}),\end{aligned}$$

where the upper and the lower bounds are constants that do not change with  $n$ .

We now check Assumptions 4 and 5. Let  $\bar{\mu}$  and  $\underline{\mu}$  be the constant upper and lower bounds for  $\dot{\mu}(X_{ij}^\top \theta)$  in the condition (2). Then for Assumption 4(i), using the boundedness of  $X_{ijk}$  in the condition (1), we have that w.p.a.1,

$$\begin{aligned}& \sup_{\|\eta\|=1} E[(\eta^\top (g(D, \theta) - g(D, \theta_0)))^2] \leq \sup_{\|\eta\|=1} E[(\eta^\top \bar{\mu} \mathbf{X}^\top \mathbf{S}(\tilde{\theta})^{-1} \bar{\mu} \mathbf{X} (\theta - \theta_0))^2] \\ & \leq \bar{\mu}^2 \underline{\lambda}(\mathbf{S}(\tilde{\theta}))^{-2} \sup_{\|\eta\|=1} \eta^\top E[\mathbf{X}^\top \mathbf{X} (\theta - \theta_0) (\theta - \theta_0)^\top \mathbf{X}^\top \mathbf{X}] \eta \\ & \leq \bar{\mu}^2 \underline{\lambda}(\mathbf{S}(\tilde{\theta}))^{-2} \bar{\lambda}(E(\mathbf{X}^\top \mathbf{X})) \text{tr}(\mathbf{X}^\top \mathbf{X}) \|\theta - \theta_0\|^2 \\ & \leq \bar{\mu}^2 \underline{\lambda}(\mathbf{S}(\tilde{\theta}))^{-2} \bar{\lambda}(E(\mathbf{X}^\top \mathbf{X})) \cdot spC_X^2 \|\theta - \theta_0\|^2 = O((p^{1/2} \|\theta - \theta_0\|)^2)\end{aligned}$$

Therefore this implies that in Assumption 4(i) we can take  $\alpha = 1$ , and also in Assumption 4(ii), the  $L_2$  norm of the envelope function  $F$  for the class  $\mathcal{F}$  is of order  $O(\sqrt{p})$ , since the  $L_2$  radius of  $\Theta$  is assumed to be bounded by constant  $R$  in Assumption 1. Next we estimate the  $L_2$  uniform covering number of  $\mathcal{F} = \{f(\eta, \theta) = \eta^\top \mathbf{B}(\tilde{\theta})^\top \mathbf{S}(\tilde{\theta})^{-1} (\mu(\theta_0) - \mu(\theta)), \theta \in \Theta, \eta \in \mathbb{R}^m, \|\eta\| = 1\}$ . Suppose there exists a  $\epsilon$ -net in  $L_2(P_D)$  norm for  $\mathcal{F} : \{(\eta_1, \theta_1), \dots, (\eta_N, \theta_N)\}$ , with  $N = N(\epsilon \|F\|_{P_D, 2}, \mathcal{F}, L_2(P_D))$ . Then by definition, for any  $(\eta, \theta)$ , one can pick out a pair  $(\eta_k, \theta_k)$ , for some  $1 \leq k \leq N$ , such that  $E|f(\eta_k, \theta_k) - f(\eta, \theta)|^2 \leq \epsilon^2$ . Then since

$$\begin{aligned}& E|f(\eta_k, \theta_k) - f(\eta, \theta)|^2 \\ & \leq 2E[(\eta_k - \eta)^\top \mathbf{B}(\tilde{\theta})^\top \mathbf{S}(\tilde{\theta})^{-1} (\mu(\theta_k) - \mu(\theta_0))]^2 + 2E[\eta^\top \mathbf{B}(\tilde{\theta})^\top \mathbf{S}(\tilde{\theta})^{-1} (\mu(\theta) - \mu(\theta_k))]^2 \\ & \leq 2\bar{\mu}^2 \underline{\lambda}(\mathbf{S}(\tilde{\theta}))^{-2} \cdot \bar{\lambda}(E(\mathbf{X}^\top \mathbf{X})) \cdot spC_X^2 \bar{\mu}^2 \cdot 4R^2 \|\eta_k - \eta\|^2 \\ & \quad + 2\|\eta\|^2 \bar{\mu}^2 \underline{\lambda}(\mathbf{S}(\tilde{\theta}))^{-2} \cdot \bar{\lambda}(E(\mathbf{X}^\top \mathbf{X})) \cdot spC_X^2 \bar{\mu}^2 \|\theta_k - \theta\|^2 \\ & \leq (C_1 p^{1/2} \|\eta_k - \eta\|)^2 + (C_2 p^{1/2} \|\theta_k - \theta\|)^2\end{aligned}$$

for some constants  $C_1, C_2 > 0$  that depend on the eigenvalues,  $R$ , and  $\bar{\mu}$ . Thus we only need  $\|\eta_k - \eta\| \leq \epsilon/(2C_1 p^{1/2})$  and  $\|\theta_k - \theta\| \leq \epsilon/(2C_2 p^{1/2})$ . Since  $\|\eta\| = 1$  and  $\|\eta_k - \eta\| \leq p^{1/2} |\eta_k - \eta|_\infty$ , we estimate the covering number on  $\eta$  using  $L_\infty$  grids and need

no more than  $N_\eta = \left(\frac{2C_1p}{\epsilon} + 1\right)^p$  points. Similarly since  $\|\theta\| \leq R$ , we need no more than  $N_\theta = \left(\frac{2C_2Rp}{\epsilon} + 1\right)^p$  points. Together we have shown that for small  $\epsilon > 0$ ,

$$N(\epsilon\|F\|_{P_D,2}, \mathcal{F}, L_2(P_D)) \leq N_\eta N_\theta \leq \left(\frac{9C_1C_2p^2R}{\epsilon^2}\right)^p,$$

which give  $\log N(\epsilon\|F\|_{P_D,2}, \mathcal{F}, L_2(P_D)) = O(p \log(n/\epsilon))$ . So Assumption 4(ii) holds.

For Assumption 5(i), we have

$$\|Eg(D, \theta)\| = \left\| E[\mathbf{B}(\tilde{\theta})^\top \mathbf{S}(\tilde{\theta})^{-1}(\mu(\theta) - \mu(\theta_0))] \right\| \geq \underline{\dot{\mu}}^2 \bar{\lambda}(\mathbf{S}(\tilde{\theta}))^{-1} \underline{\lambda}(E[\mathbf{X}^\top \mathbf{X}]) \|\theta - \theta_0\|.$$

Therefore Assumption 5(i) holds with  $\delta_1 = \underline{\dot{\mu}}^2 \bar{\lambda}(\mathbf{S}(\tilde{\theta}))^{-1} \underline{\lambda}(E[\mathbf{X}^\top \mathbf{X}])$  and  $\delta_0 = R\delta_1$ .

For Assumption 5(ii),  $\mathbf{G} = \nabla_\theta Eg(D, \theta_0) = -E[\mathbf{B}(\tilde{\theta})^\top \mathbf{S}(\tilde{\theta})^{-1} \mathbf{B}(\theta_0)]$ . By conditions (2) and (3),

$$\begin{aligned} \bar{\lambda}(\mathbf{G}^\top \mathbf{G}) &\leq \bar{\mu}^4 \underline{\lambda}(\mathbf{S}(\tilde{\theta}))^{-2} \bar{\lambda}(E[\mathbf{X}^\top \mathbf{X}])^2 \\ \underline{\lambda}(\mathbf{G}^\top \mathbf{G}) &\leq \underline{\dot{\mu}}^4 \bar{\lambda}(\mathbf{S}(\tilde{\theta}))^{-2} \underline{\lambda}(E[\mathbf{X}^\top \mathbf{X}])^2 \end{aligned}$$

so the eigenvalues of  $\mathbf{G}^\top \mathbf{G}$  are bounded above and below as  $n \rightarrow \infty$ .

In Assumption 5(iii), let  $\mathbf{K}(\theta) = (\ddot{\mu}(X_{\cdot 1}^\top \theta)(X_{\cdot 1} \otimes X_{\cdot 1}), \dots, \ddot{\mu}(X_{\cdot s}^\top \theta)(X_{\cdot s} \otimes X_{\cdot s}))^\top$ . Then for any unit vectors  $u, v \in \mathbb{R}^p$ ,

$$\|\mathbf{H}(\theta)[u, v]\| = \|E[\mathbf{B}(\tilde{\theta})^\top \mathbf{S}(\tilde{\theta})^{-1} \mathbf{K}(\theta)] \cdot \text{vec}(u \otimes v)\| \leq \bar{\mu} \bar{\mu} \bar{\lambda}(\mathbf{S}(\tilde{\theta}))^{-1} \bar{\lambda}(E[\mathbf{X}^\top \mathbf{X}]) \cdot \sqrt{p} C_X,$$

where we used the upper bound on  $\ddot{\mu}(X_{ij}^\top \theta)$  for any  $\theta \in B_0(c\epsilon_n)$  in condition (3). Therefore Assumption 5(iii) holds.

To show Assumption 6, we note that since  $\mathbf{V}_n$  and  $\mathbf{V}$  are symmetric positive definite matrices,  $\|\mathbf{V}\| = \sqrt{\bar{\lambda}(\mathbf{V}^\top \mathbf{V})} = \bar{\lambda}(\mathbf{V})$  and also  $\|\mathbf{V}_n\| = \bar{\lambda}(\mathbf{V}_n)$ . If we can show  $\|\mathbf{V}_n - \mathbf{V}\| \rightarrow 0$  w.p.a.1 and the eigenvalues of  $\mathbf{V}$  are bounded from above and below, then we have

$$\begin{aligned} \bar{\lambda}(\mathbf{V}_n) &= \|\mathbf{V}_n\| \leq \|\mathbf{V}_n - \mathbf{V}\| + \|\mathbf{V}\| = \|\mathbf{V}_n - \mathbf{V}\| + \bar{\lambda}(\mathbf{V}) \\ \underline{\lambda}(\mathbf{V}_n) &= \min_{\eta \in \mathbb{R}^p} \eta^\top \mathbf{V}_n \eta \geq \min_{\eta \in \mathbb{R}^p} \eta^\top \mathbf{V} \eta - \max_{\eta \in \mathbb{R}^p} \eta^\top (\mathbf{V} - \mathbf{V}_n) \eta \\ &\geq \underline{\lambda}(\mathbf{V}) - \|\mathbf{V}_n - \mathbf{V}\|. \end{aligned}$$

Therefore w.p.a.1, the eigenvalues of  $\mathbf{V}_n$  are also bounded from above and below, as long as  $\|\mathbf{V}_n - \mathbf{V}\| \rightarrow 0$ . Next we show the boundedness for the eigenvalues of  $\mathbf{V}$  and the convergence of  $\|\mathbf{V}_n - \mathbf{V}\|$ , respectively.

Since  $Eg(D, \theta_0) = 0$ , we have

$$\begin{aligned} \mathbf{V} &= \text{Var}(g(D, \theta_0)) = E \left[ \mathbf{B}(\theta_0)^\top \mathbf{S}(\theta_0)^{-1} (Y - \mu(\theta_0)) (Y - \mu(\theta_0))^\top \mathbf{S}(\theta_0)^{-1} \mathbf{B}(\theta_0) \right] \\ &= E \left[ \mathbf{B}(\theta_0)^\top \mathbf{S}(\theta_0)^{-1} \mathbf{S}(\theta_0) \mathbf{S}(\theta_0)^{-1} \mathbf{B}(\theta_0) \right] \\ &= E \left[ \mathbf{B}(\theta_0)^\top \mathbf{S}(\theta_0)^{-1} \mathbf{B}(\theta_0) \right]. \end{aligned}$$

By a similar argument to the boundedness of eigenvalues of  $\mathbf{S}(\tilde{\theta})$ , one can show that the eigenvalues of  $\mathbf{S}(\theta_0)$  are also bounded from above and below by constants. Therefore,

$$\begin{aligned} \bar{\lambda}(\mathbf{V}) &\leq \bar{\lambda}(E(\mathbf{B}(\theta_0)^\top \mathbf{B}(\theta_0))) \underline{\lambda}(\mathbf{S}(\theta_0))^{-1} \\ &\leq \bar{\mu}^2 \bar{\lambda}(E(\mathbf{X}^\top \mathbf{X})) \underline{\lambda}(\mathbf{S}(\theta_0))^{-1} \\ \underline{\lambda}(\mathbf{V}) &\geq \underline{\lambda}(E(\mathbf{B}(\theta_0)^\top \mathbf{B}(\theta_0))) \bar{\lambda}(\mathbf{S}(\theta_0))^{-1} \\ &\geq \underline{\mu}^2 \underline{\lambda}(E(\mathbf{X}^\top \mathbf{X})) \bar{\lambda}(\mathbf{S}(\theta_0))^{-1}. \end{aligned}$$

The boundedness of  $\bar{\lambda}(\mathbf{V})$  and  $\underline{\lambda}(\mathbf{V})$  is proved.

To show  $\|\mathbf{V}_n - \mathbf{V}\| \rightarrow 0$ , we first note that

$$\begin{aligned} \mathbf{V}_n &= \frac{1}{n} \sum_{i=1}^n (g(D_i, \tilde{\theta}) - \bar{g}(\mathbf{D}, \tilde{\theta})) (g(D_i, \tilde{\theta}) - \bar{g}(\mathbf{D}, \tilde{\theta}))^\top \\ &= \frac{1}{n} \sum_{i=1}^n g(D_i, \tilde{\theta}) g(D_i, \tilde{\theta})^\top + \frac{1}{n} \sum_{i=1}^n \bar{g}(\mathbf{D}, \tilde{\theta}) \bar{g}(\mathbf{D}, \tilde{\theta})^\top \\ &= \frac{1}{n} \sum_{i=1}^n g(D_i, \theta_0) g(D_i, \theta_0)^\top + \frac{2}{n} \sum_{i=1}^n (g(D_i, \tilde{\theta}) - g(D_i, \theta_0)) g(D_i, \theta_0)^\top \\ &\quad + \frac{1}{n} \sum_{i=1}^n (g(D_i, \tilde{\theta}) - g(D_i, \theta_0)) (g(D_i, \tilde{\theta}) - g(D_i, \theta_0))^\top + \bar{g}(\mathbf{D}, \tilde{\theta}) \bar{g}(\mathbf{D}, \tilde{\theta})^\top \\ &:= \mathbf{E}_1 + \mathbf{E}_2 + \mathbf{E}_3 + \mathbf{E}_4. \end{aligned} \tag{A.17}$$

We derive bounds for each term. For  $\mathbf{E}_1$  we have

$$\begin{aligned} \|\mathbf{E}_1 - \mathbf{V}\|^2 &\leq \|\mathbf{E}_1 - \mathbf{V}\|_F^2 \\ &= \sum_{j=1}^p \sum_{k=1}^p \left[ n^{-1} \sum_{i=1}^n g_j(D_i, \theta_0) g_k(D_i, \theta_0) - E[g_j(D, \theta_0) g_k(D, \theta_0)] \right]^2 \end{aligned}$$

Hence by Chebyshev's inequality, for any  $C > 0$ ,

$$\begin{aligned} P(\|\mathbf{E}_1 - \mathbf{V}\|^2 > C) &\leq C^{-2} \sum_{j=1}^p \sum_{k=1}^p E \left[ n^{-1} \sum_{i=1}^n g_j(D_i, \theta_0) g_k(D_i, \theta_0) - E[g_j(D, \theta_0) g_k(D, \theta_0)] \right]^2 \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{nC^2} \sum_{j=1}^p \sum_{k=1}^p \text{Var}(g_j(D, \theta_0)g_k(D, \theta_0)) \\
&\leq \frac{p^2}{nC^2} \sup_{1 \leq j, k \leq p} E[g_j(D, \theta_0)^2 g_k(D, \theta_0)^2] \leq \frac{p^2}{nC^2} \sup_{1 \leq j \leq p} E[g_j(D, \theta_0)^4] \\
&\leq \frac{p^2}{nC^2} \sup_{1 \leq j \leq p} \bar{\mu}^4 E[X_{j\cdot}^\top \mathbf{S}(\theta_0)^{-1} (Y - \mu(\theta_0))]^4 \\
&\leq \frac{p^2}{nC^2} \bar{\mu}^4 s^2 C_X^4 \underline{\lambda}(\mathbf{S}(\theta_0))^{-4} E\|Y - \mu(\theta_0)\|^4 \\
&= \frac{p^2}{nC^2} \bar{\mu}^4 s^2 C_X^4 \underline{\lambda}(\mathbf{S}(\theta_0))^{-4} E\left[\sum_{j=1}^s (Y_j - \mu_j(\theta_0))^2\right]^2 \\
&\leq \frac{p^2}{nC^2} \bar{\mu}^4 s^3 C_X^4 \underline{\lambda}(\mathbf{S}(\theta_0))^{-4} E\left[\sum_{j=1}^s (Y_j - \mu_j(\theta_0))^4\right] \\
&\leq \frac{8p^2}{nC^2} \bar{\mu}^4 s^3 C_X^4 \underline{\lambda}(\mathbf{S}(\theta_0))^{-4} E\sum_{j=1}^s [Y_j^4 + \mu_j(\theta_0)^4] \\
&\leq \frac{16p^2}{nC^2} \bar{\mu}^4 s^4 C_X^4 \underline{\lambda}(\mathbf{S}(\theta_0))^{-4} \sup_{1 \leq j \leq s} E(Y_j^4).
\end{aligned}$$

Since  $\sup_{1 \leq j \leq s} E(Y_j^4) < \infty$  as in Condition (1), we conclude that  $\|\mathbf{E}_1 - \mathbf{V}\| = O_p(p/\sqrt{n}) = o_p(1)$ .

Next we bound  $\mathbf{E}_3$ . Because  $\|\tilde{\theta} - \theta_0\| = O_p(\sqrt{p/n})$ , we have that for any generic  $D$ ,

$$\begin{aligned}
&\|g(D, \tilde{\theta}) - g(D, \theta_0)\| = \|\mathbf{B}(\tilde{\theta})^\top \mathbf{S}(\tilde{\theta})^{-1} (\mu(\tilde{\theta}) - \mu(\theta_0))\| \\
&= \|\mathbf{B}(\tilde{\theta})^\top \mathbf{S}(\tilde{\theta})^{-1} \mathbf{B}(\theta') (\tilde{\theta} - \theta_0)\| \leq \|\mathbf{B}(\tilde{\theta})^\top \mathbf{S}(\tilde{\theta})^{-1} \mathbf{B}(\theta')\| \cdot \|\tilde{\theta} - \theta_0\| \\
&\leq \bar{\mu}^2 \underline{\lambda}(\mathbf{S}(\tilde{\theta})) \cdot \|\mathbf{X}^\top \mathbf{X}\| \cdot \|\tilde{\theta} - \theta_0\| \leq \bar{\mu}^2 \underline{\lambda}(\mathbf{S}(\tilde{\theta})) \cdot \|\mathbf{X}\|^2 \cdot O_p\left(\sqrt{\frac{p}{n}}\right) \\
&\leq \bar{\mu}^2 \underline{\lambda}(\mathbf{S}(\tilde{\theta})) \cdot sp C_X^2 \cdot O_p\left(\sqrt{\frac{p}{n}}\right) = O_p\left(\sqrt{\frac{p^3}{n}}\right) = o_p(1),
\end{aligned} \tag{A.18}$$

where  $\theta'$  is between  $\tilde{\theta}$  and  $\theta_0$ , and the derivation shows that if we replace  $\mathbf{D}$  with  $\mathbf{D}_i$ , then the upper bound is uniform over all  $i = 1, \dots, n$ . Therefore

$$\begin{aligned}
\|\mathbf{E}_3\| &\leq n^{-1} \sum_{i=1}^n \left\| (g(D_i, \tilde{\theta}) - g(D_i, \theta_0)) (g(D_i, \tilde{\theta}) - g(D_i, \theta_0))^\top \right\| \\
&\leq n^{-1} \sum_{i=1}^n \left\| g(D_i, \tilde{\theta}) - g(D_i, \theta_0) \right\|^2 = O_p\left(\frac{p^3}{n}\right).
\end{aligned}$$

Given the bounds for  $\mathbf{E}_1$  and  $\mathbf{E}_3$ , we can bound  $\mathbf{E}_2$  as

$$\begin{aligned}
\|\mathbf{E}_2\| &= 2n^{-1} \left\| \sum_{i=1}^n (g(D_i, \tilde{\theta}) - g(D_i, \theta_0)) g(D_i, \theta_0)^\top \right\| \\
&\leq 2n^{-1} \sum_{i=1}^n \left\| g(D_i, \tilde{\theta}) - g(D_i, \theta_0) \right\| \cdot \left\| g(D_i, \theta_0) \right\|
\end{aligned}$$

$$\begin{aligned}
&\leq O_p\left(\sqrt{\frac{p^3}{n}}\right) \cdot 2n^{-1} \sum_{i=1}^n \left\|g(D_i, \theta_0)\right\| \\
&\leq O_p\left(\sqrt{\frac{p^3}{n}}\right) \cdot O_p\left(\sqrt{\frac{p}{n}}\right) = O_p\left(\frac{p^2}{n}\right) = o_p(1).
\end{aligned}$$

For  $\mathbf{E}_4$ , we use  $\|\bar{g}(\mathbf{D}, \theta_0)\| = O_p(\sqrt{p/n})$  and (A.18)

$$\begin{aligned}
\|\mathbf{E}_4\| &= \|\bar{g}(\mathbf{D}, \tilde{\theta})\|^2 \leq (\|\bar{g}(\mathbf{D}, \tilde{\theta}) - \bar{g}(\mathbf{D}, \theta_0)\| + \|\bar{g}(\mathbf{D}, \theta_0)\|)^2 \\
&\leq (O_p(\sqrt{p^3/n}) + O_p(\sqrt{p/n}))^2 = O_p(p^3/n) = o_p(1).
\end{aligned}$$

Finally, we combine the bounds for  $\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3, \mathbf{E}_4$  and conclude that  $\|\mathbf{V}_n - \mathbf{V}\| = o_p(1)$ . Therefore Assumption 6 holds.  $\blacksquare$

### Proof of Theorem 3:

We check Assumptions 4 and 5. For a generic  $\theta$ , let  $A = \{Y \text{ is between } X^\top \theta \text{ and } X^\top \theta_0\}$ . Let  $\bar{f}$ ,  $\tilde{f}$  and  $\underline{f}$  be the upper bounds for the conditional density  $f_{Y|X}$ , its derivative  $\dot{f}_{Y|X}$  and the lower bound for  $f_{Y|X}$  in condition (2). Then

$$\begin{aligned}
&\sup_{\|\eta\|=1} E[\eta^\top (g(D, \theta) - g(D, \theta_0))]^2 = \sup_{\|\eta\|=1} E[\eta^\top X(1(Y \leq X^\top \theta) - 1(Y \leq X^\top \theta_0))]^2 \\
&= \sup_{\|\eta\|=1} \eta^\top E_X \left[ X \left( \int_A f_{Y|X}(y) dy \right) X^\top \right] \eta \\
&\leq \bar{\lambda}(E[XX^\top]) \bar{f} C_X p^{1/2} \|\theta - \theta_0\|.
\end{aligned}$$

Therefore Assumption 4(i) follows by taking  $\alpha = 1/2$ , since the eigenvalues of  $E[XX^\top]$  are bounded by condition (3). It also implies that the  $L_2$  norm for the envelope function  $F$  of the class  $\mathcal{F}$  in Assumption 4(ii) is bounded by  $O(R^{1/2} p^{1/4}) \leq O(p^{1/2})$ . Moreover, the VC index of the class  $\mathcal{F}$  is of order  $O(p)$  (see Lemma 18-20 of Belloni et al. 2011), and the bound on the uniform covering number follows by Theorem 2.6.7 of van der Vaart and Wellner (1996).

For Assumption 5(i), we have

$$\begin{aligned}
\|Eg(D, \theta)\|^2 &= \|E[X(1(Y \leq X^\top \theta) - \tau)]\|^2 = \|E[X(F_{Y|X}(X^\top \theta) - F_{Y|X}(X^\top \theta_0))]\|^2 \\
&= \|E[XX^\top f_{Y|X}(X^\top \tilde{\theta}) \cdot (\theta - \theta_0)]\|^2 \geq \underline{f}^2 \underline{\lambda}(E[XX^\top]) \|\theta - \theta_0\|^2,
\end{aligned}$$

where in the second equality we used the iterated expectation, in the third equality  $\tilde{\theta}$  is between  $\theta$  and  $\theta_0$ . This implies that  $\|Eg(D, \theta)\| \geq \delta_1 \|\theta - \theta_0\|$ , with  $\delta_1 = \underline{f}^2 \underline{\lambda}(E[XX^\top])$ . Therefore we can simply take  $\delta_0 = 2R\delta_1$ , and Assumption 5(i) holds.

For Assumption 5(ii), one can calculate that  $\mathbf{G} = E[XX^\top f_{Y|X}(X^\top \theta_0)]$ . Using the definition of the matrix operator norm, one can see that the eigenvalues of  $\mathbf{G}^\top \mathbf{G}$  can be bounded as

$$\begin{aligned}\bar{\lambda}(\mathbf{G}^\top \mathbf{G}) &\leq \bar{f}^2 \bar{\lambda}(E(XX^\top)) \\ \underline{\lambda}(\mathbf{G}^\top \mathbf{G}) &\geq \underline{f}^2 \underline{\lambda}(E(XX^\top))\end{aligned}$$

For Assumption 5(iii), for any unit vectors  $u, v \in \mathbb{R}^p$ ,

$$\begin{aligned}\|\mathbf{H}(\theta)[u, v]\| &= \|E[XX^\top \otimes X^\top \dot{f}_{Y|X}(X^\top \theta)] \cdot \text{vec}(u \otimes v)\| \\ &\leq \bar{\lambda}(E[XX^\top]) \cdot \sqrt{p} C_X \bar{f}.\end{aligned}$$

Hence Assumption 5(iii) holds.

For Assumption 6, by Chebyshev's inequality, for any  $C > 0$ , we have

$$\begin{aligned}P(\|\mathbf{V}_n - \mathbf{V}\| \geq C) &\leq P(\|\mathbf{V}_n - \mathbf{V}\|_F \geq C) \\ &\leq \frac{\tau^2(1-\tau)^2}{C^2} \text{Var} \left[ \sum_{j=1}^p \sum_{k=1}^p \left( \frac{1}{n} \sum_{i=1}^n X_{ij} X_{ik} - E(X_{ij} X_{ik}) \right)^2 \right] \\ &\leq \frac{\tau^2(1-\tau)^2 p^2}{nC^2} \sup_{1 \leq j, k \leq p} E[X_j^2 X_k^2] \leq \frac{p^2 C_X^4 \tau^2 (1-\tau)^2}{nC^2}.\end{aligned}$$

Therefore  $\|\mathbf{V}_n - \mathbf{V}\| = O_p(p/\sqrt{n}) = o_p(1)$ . The boundedness of eigenvalues of  $\mathbf{V}$  follows directly from the boundedness of eigenvalues of  $E[XX^\top]$  in Condition (3), and hence the eigenvalues of  $\mathbf{V}_n$  are also bounded from above and below w.p.a.1.  $\blacksquare$

#### Proof of Theorem 4:

Hereafter we denote the  $(i, j)$ th entry in a generic  $s \times s$  positive definite matrix  $\mathbf{\Sigma}$  or  $\mathbf{\Omega}$  as  $\sigma_{ij}$  or  $\omega_{ij}$ , respectively. Denote the  $(i, j)$ th entry in the true covariance matrix  $\mathbf{\Sigma}_0$  and the true precision matrix  $\mathbf{\Omega}_0$  as  $\sigma_{ij,0}$  or  $\omega_{ij,0}$ , respectively. For a generic parameter  $\theta$ , we denote the corresponding precision matrix as  $\mathbf{\Omega}$  and the corresponding covariance matrix as  $\mathbf{\Sigma} = \mathbf{\Omega}^{-1}$ . The coordinates of  $\theta$  and any other  $p$ -dimensional vector is subscripted by “ $ij$ ” with  $1 \leq i \leq j \leq s$ . Then we can first establish an equivalence between the  $L_2$  norm of  $\theta$  and the Frobenius norm of  $\mathbf{\Omega}$ . Since  $\theta$  contains the entries in the upper triangle of  $\mathbf{\Omega}$ , it is obvious that

$$\frac{1}{2} \|\mathbf{\Omega} - \mathbf{\Omega}_0\|_F^2 \leq \|\theta - \theta_0\|^2 \leq \|\mathbf{\Omega} - \mathbf{\Omega}_0\|_F^2, \quad (\text{A.19})$$

so these two norms are equivalent.

Now we check Assumptions 4 and 5. For Assumption 4(i), since  $m = p = s(s+1)/2$  for this example, we have that for any  $\eta \in \mathbb{R}^p$  and  $\|\eta\|^2 = \sum_{1 \leq i \leq j \leq s} \eta_{ij}^2 = 1$ , by the Cauchy-Schwarz inequality,

$$\begin{aligned}
E[\eta^\top (g(D, \theta) - g(D, \theta_0))]^2 &= \left[ \sum_{1 \leq i \leq j \leq s} \eta_{ij} (\sigma_{ij} - \sigma_{ij,0}) \right]^2 \\
&\leq \sum_{1 \leq i \leq j \leq s} \eta_{ij}^2 \cdot \sum_{1 \leq i \leq j \leq s} (\sigma_{ij} - \sigma_{ij,0})^2 \leq \sum_{1 \leq i \leq s, 1 \leq j \leq s} (\sigma_{ij} - \sigma_{ij,0})^2 \\
&= \|\Sigma - \Sigma_0\|_F^2 = \|\Omega^{-1}(\Omega_0 - \Omega)\Omega_0^{-1}\|_F^2 \leq \|\Omega^{-1}\|_F^2 \|\Omega - \Omega_0\|_F^2 \|\Omega_0^{-1}\|_F^2 \\
&\leq \underline{\lambda}(\Omega)^{-2} \underline{\lambda}(\Omega_0)^{-2} s^2 \|\Omega - \Omega_0\|_F^2 = O((p^{1/2} \|\theta - \theta_0\|)^2), \tag{A.20}
\end{aligned}$$

where we used the submultiplicativity of the Frobenius norm, the boundedness of eigenvalues in the condition (1), the relation  $\|\mathbf{A}\|_F^2 \leq s \bar{\lambda}(\mathbf{A})^2$  for a  $s \times s$  positive definite matrix  $\mathbf{A}$ , the relation  $p = s(s+1)/2$  and (A.19). Take supremum over  $\eta$  and Assumption 4(i) is proved.

For Assumption 4(ii), we have derived above that the envelope function of  $\mathcal{F}$  has  $L_2$  norm of order  $O(p^{1/2})$  given  $\|\theta\| \leq R$ . Note that in fact for the partial correlation selection example, the functions in  $\mathcal{F}$  do not have any randomness. Suppose a  $L_2$   $\epsilon$ -net of  $\mathcal{F}$  is  $\{(\eta_1, \theta_1), \dots, (\eta_N, \theta_N)\}$  for  $N = N(\epsilon \|F\|_{P_D, 2}, \mathcal{F}, L_2(P_D))$ . Then for any  $f(\eta, \theta) \in \mathcal{F}$ , we apply a similar procedure of (A.20) and have

$$\begin{aligned}
E|f(\eta_k, \theta_k) - f(\eta, \theta)|^2 &\leq 2|f(\eta_k, \theta_k) - f(\eta, \theta_k)|^2 + 2|f(\eta, \theta_k) - f(\eta, \theta)|^2 \\
&\leq 2\|\eta_k - \eta\|^2 \underline{\lambda}(\Omega_k)^{-2} \underline{\lambda}(\Omega_0)^{-2} s^2 \cdot 4R^2 + 2\underline{\lambda}(\Omega_k)^{-2} \underline{\lambda}(\Omega)^{-2} s^2 \cdot 2\|\theta_k - \theta\|^2 \\
&:= (C_1 p^{1/2} \|\eta_k - \eta\|)^2 + (C_2 p^{1/2} \|\theta_k - \theta\|)^2,
\end{aligned}$$

where  $\Omega_k$  is the matrix  $\Omega$  with parameter  $\theta_k$ . Therefore by a similar argument to the proof of Theorem 2,  $N(\epsilon \|F\|_{P_D, 2}, \mathcal{F}, L_2(P_D)) \leq \left( \frac{9C_1 C_2 p^2 R}{\epsilon^2} \right)^p$ , which give  $\log N(\epsilon \|F\|_{P_D, 2}, \mathcal{F}, L_2(P_D)) = O(p \log(n/\epsilon))$  and hence Assumption 4(ii) holds.

For Assumption 5(i), we have

$$\begin{aligned}
\|Eg(D, \theta)\|^2 &= \sum_{1 \leq i \leq j \leq s} (\sigma_{ij} - \sigma_{ij,0})^2 \geq \frac{1}{2} \sum_{1 \leq i \leq s, 1 \leq j \leq s} (\sigma_{ij} - \sigma_{ij,0})^2 \\
&= \frac{1}{2} \|\Sigma - \Sigma_0\|_F^2 = \frac{1}{2} \|\Omega^{-1}(\Omega_0 - \Omega)\Omega_0^{-1}\|_F^2 \geq \frac{1}{2} \bar{\lambda}(\Omega)^{-2} \bar{\lambda}(\Omega_0)^{-2} \|\Omega - \Omega_0\|_F^2, \tag{A.21}
\end{aligned}$$

where we have used the fact that for two positive definite matrices  $\mathbf{A}$  and  $\mathbf{B}$ ,

$$\|\mathbf{AB}\|_F^2 = \text{tr}(\mathbf{B}^\top \mathbf{A}^\top \mathbf{AB}) \geq \underline{\lambda}(\mathbf{B})^2 \text{tr}(\mathbf{A}^\top \mathbf{A}) \geq \underline{\lambda}(\mathbf{B})^2 \|\mathbf{A}\|_F^2.$$

Now since we have assumed in the condition (1) that the eigenvalues of  $\mathbf{\Omega}$  are bounded above by constants, (A.21) implies that  $\|Eg(D, \theta)\| \geq \delta_1 \|\theta - \theta_0\|$  with  $0 < \delta_1 < \bar{\lambda}(\mathbf{\Omega})^{-1} \bar{\lambda}(\mathbf{\Omega}_0)^{-1} / \sqrt{2}$ . So Assumption 5(i) holds with this  $\delta_1$  and  $\delta_0 = R\delta_1$ .

To show Assumption 5(ii), we only need to show that for any unit vector  $u \in \mathbb{R}^p$ ,  $u^\top \mathbf{G}^\top \mathbf{G} u$  is bounded above and below by constants, where  $\mathbf{G} = \nabla_\theta Eg(D, \theta_0)$ . Define a linear operator  $\partial_u := \sum_{1 \leq i \leq j \leq s} u_{ij} \frac{\partial}{\partial \theta_{ij}}$  and define  $u_{ji} := u_{ij}$  for any  $j < i$ . Then  $u^\top \mathbf{G}^\top \mathbf{G} u = \partial_u Eg(D, \theta_0)^\top \partial_u Eg(D, \theta_0) = \|\partial_u Eg(D, \theta_0)\|^2$ , and similar to (A.19), one can show that

$$\frac{1}{2} \|\partial_u \mathbf{\Omega}_0^{-1}\|_F^2 \leq \|\partial_u Eg(D, \theta_0)\|^2 \leq \|\partial_u \mathbf{\Omega}_0^{-1}\|_F^2,$$

and also

$$1 = \|u\|^2 \leq \|\partial_u \mathbf{\Omega}_0\|_F^2 \leq 2\|u\|^2 = 2.$$

Since  $\mathbf{\Omega} \mathbf{\Omega}^{-1} = \mathbf{I}$ , we take first derivative and have  $\partial_u \mathbf{\Omega}^{-1} = -\mathbf{\Omega}^{-1}(\partial_u \mathbf{\Omega}) \mathbf{\Omega}^{-1}$ . Therefore, we have

$$\begin{aligned} \|\partial_u \mathbf{\Omega}_0^{-1}\|_F^2 &= \|\mathbf{\Omega}_0^{-1}(\partial_u \mathbf{\Omega}_0) \mathbf{\Omega}_0^{-1}\|_F^2 \leq \underline{\lambda}(\mathbf{\Omega}_0)^{-4} \|\partial_u \mathbf{\Omega}_0\|_F^2 \leq 2\underline{\lambda}(\mathbf{\Omega}_0)^{-4} \\ \|\partial_u \mathbf{\Omega}_0^{-1}\|_F^2 &= \|\mathbf{\Omega}_0^{-1}(\partial_u \mathbf{\Omega}_0) \mathbf{\Omega}_0^{-1}\|_F^2 \geq \bar{\lambda}(\mathbf{\Omega}_0)^{-4} \|\partial_u \mathbf{\Omega}_0\|_F^2 \geq \bar{\lambda}(\mathbf{\Omega}_0)^{-4}, \end{aligned}$$

which implies the boundedness of eigenvalues of  $\mathbf{G}^\top \mathbf{G}$ , given the condition (1).

For Assumption 5(iii), we use the same technique and have that for unit vectors  $u, v \in \mathbb{R}^p$ ,  $\|\mathbf{H}(\theta)(u, v)\|^2 = \|\partial_u \partial_v Eg(D, \theta)\|^2 \leq \|\partial_u \partial_v \mathbf{\Omega}^{-1}\|_F^2$ . While for any generic  $\mathbf{\Omega}$ , using  $\partial_u \mathbf{\Omega}^{-1} = -\mathbf{\Omega}^{-1}(\partial_u \mathbf{\Omega}) \mathbf{\Omega}^{-1}$ , we get

$$\partial_u \partial_v \mathbf{\Omega}^{-1} = -\mathbf{\Omega}^{-1}(\partial_u \mathbf{\Omega}) \mathbf{\Omega}^{-1}(\partial_v \mathbf{\Omega}) \mathbf{\Omega}^{-1} - \mathbf{\Omega}^{-1}(\partial_v \mathbf{\Omega}) \mathbf{\Omega}^{-1}(\partial_u \mathbf{\Omega}) \mathbf{\Omega}^{-1} - \mathbf{\Omega}^{-1}(\partial_u \partial_v \mathbf{\Omega}) \mathbf{\Omega}^{-1}.$$

Therefore since  $\partial_u \partial_v \mathbf{\Omega} \equiv 0$  for any  $\mathbf{\Omega}$ ,

$$\begin{aligned} \|\mathbf{H}(\theta)(u, v)\|^2 &\leq \|\mathbf{\Omega}^{-1}(\partial_u \mathbf{\Omega}) \mathbf{\Omega}^{-1}(\partial_v \mathbf{\Omega}) \mathbf{\Omega}^{-1}\|_F^2 + \|\mathbf{\Omega}^{-1}(\partial_v \mathbf{\Omega}) \mathbf{\Omega}^{-1}(\partial_u \mathbf{\Omega}) \mathbf{\Omega}^{-1}\|_F^2 \\ &\leq 2\underline{\lambda}(\mathbf{\Omega})^{-6} + 2\underline{\lambda}(\mathbf{\Omega})^{-6} = 4\underline{\lambda}(\mathbf{\Omega})^{-6}, \end{aligned}$$

which is bounded above by constant for any  $\mathbf{\Omega}$  considered here by the condition (1).

Hence Assumption 5(iii) holds.

For Assumption 6, we have assumed the boundedness of eigenvalues of  $\mathbf{V}$ , and we still need to show the convergence  $\|\mathbf{V}_n - \mathbf{V}\| \rightarrow 0$  w.p.a.1. By Chebyshev's inequality, for any  $C > 0$ , we have

$$P(\|\mathbf{V}_n - \mathbf{V}\| \geq C) \leq P(\|\mathbf{V}_n - \mathbf{V}\|_F \geq C)$$



$$\begin{aligned}
&\leq \frac{1}{C^2} \sum_{1 \leq j_1 \leq k_1 \leq s} \sum_{1 \leq j_2 \leq k_2 \leq s} \text{Var} \left[ \frac{1}{n} \sum_{i=1}^n (Y_{ij_1} Y_{ik_1} - \sigma_{j_1 k_1, 0})(Y_{ij_2} Y_{ik_2} - \sigma_{j_2 k_2, 0}) \right. \\
&\quad \left. - E[(Y_{j_1} Y_{k_1} - \sigma_{j_1 k_1, 0})(Y_{j_2} Y_{k_2} - \sigma_{j_2 k_2, 0})] \right] \\
&\leq \frac{s^2(s+1)^2}{4nC^2} \sup_{j_1, k_1, j_2, k_2} E \left[ (Y_{j_1} Y_{k_1} - \sigma_{j_1 k_1, 0})^2 (Y_{j_2} Y_{k_2} - \sigma_{j_2 k_2, 0})^2 \right] \\
&\leq \frac{4s^2(s+1)^2}{nC^2} \sup_{j_1, k_1, j_2, k_2} E[Y_{j_1}^2 Y_{k_1}^2 Y_{j_2}^2 Y_{k_2}^2] \leq \frac{16p^2}{nC^2} \sup_{1 \leq j \leq s} E(Y_j^8).
\end{aligned}$$

Hence we have  $\|\mathbf{V}_n - \mathbf{V}\| = O_p(p/\sqrt{n}) = o_p(1)$ . This together with Condition (3) implies the boundedness of eigenvalues of  $\mathbf{V}_n$  w.p.a.1. Therefore Assumption 6 holds.  $\blacksquare$

## Appendix B. Proof for the asymptotic validity of the BGMM inference

In this appendix, we give the proofs of Theorem 5 and Theorem 6. For the ease of notation, let  $\Sigma(\theta) = \mathbf{G}(\theta)^\top \mathbf{V}(\theta)^{-1} \mathbf{G}(\theta)$ , where  $\mathbf{G}(\theta)$  and  $\mathbf{V}(\theta)$  are defined in Assumption 10. In the following, for any matrix  $\mathbf{A}(\theta)$  that depends on  $\theta$ , we use the notation “ $\mathbf{A}$ ” to refer to the matrix evaluated at  $\theta_0$ . For example,  $\Sigma = \mathbf{G}^\top \mathbf{V}^{-1} \mathbf{G}$ , where  $\mathbf{G}$  and  $\mathbf{V}$  are defined in Assumption 5 and 6, i.e.  $\mathbf{G}(\theta)$  and  $\mathbf{V}(\theta)$  evaluated at  $\theta = \theta_0$ , respectively. For any model  $\mathcal{M}$ , one can partition any matrix  $\mathbf{G}(\theta)$  into  $\mathbf{G}(\theta) = (\mathbf{G}_{\mathcal{M}}(\theta), \mathbf{G}_{\mathcal{M}^c}(\theta))$ , according to the partial derivative with respect to the components of  $\theta$  in either  $\mathcal{M}$  or  $\mathcal{M}^c$ . Let

$$\begin{aligned}\Sigma_{11}(\theta) &= \mathbf{G}_{\mathcal{M}}(\theta)^\top \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}}(\theta) \\ \Sigma_{12}(\theta) &= \mathbf{G}_{\mathcal{M}}(\theta)^\top \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}^c}(\theta) \\ \Sigma_{22}(\theta) &= \mathbf{G}_{\mathcal{M}^c}(\theta)^\top \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}^c}(\theta) \\ \mathbf{J}_{\mathcal{M}}(\theta) &= \mathbf{I}_p - \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}}(\theta) (\mathbf{G}_{\mathcal{M}}(\theta)^\top \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}}(\theta))^{-1} \mathbf{G}_{\mathcal{M}}(\theta)^\top\end{aligned}$$

We have the following lemma about the quadratic term in the asymptotic normal density of the GMM estimator  $\hat{\theta}$ . The proof is straightforward algebra.

**Lemma A.7.** *Under Assumptions 1-10,*

$$\begin{aligned}\frac{n}{2}(\theta - \hat{\theta})^\top \Sigma(\theta)(\theta - \hat{\theta}) &= \frac{n}{2}(\theta_1 - \xi_1(\theta))^\top \Sigma_{11}(\theta)(\theta_1 - \xi_1(\theta)) + T_{\mathcal{M}}(\theta) \\ \xi_1(\theta) &:= \hat{\theta}_1 + \Sigma_{11}(\theta)^{-1} \Sigma_{12}(\theta) \hat{\theta}_2 \\ T_{\mathcal{M}}(\theta) &:= \frac{n}{2} \hat{\theta}_2^\top \mathbf{G}_{\mathcal{M}^c}(\theta)^\top \mathbf{J}_{\mathcal{M}}(\theta) \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}^c}(\theta) \hat{\theta}_2\end{aligned}$$

where  $\hat{\theta} = (\hat{\theta}_1^\top, \hat{\theta}_2^\top)^\top$  is the GMM estimator on the full model space in  $\mathbb{R}^p$ .

**Lemma A.8.** *Suppose Assumptions 1-10 hold. Then uniformly for all  $\theta \in B_0(C\epsilon_n) \cap \Theta(\mathcal{M})$  and all  $\mathcal{M} \supseteq \mathcal{M}_0$ , with any fixed constant  $C > 0$ , for  $\xi_1(\theta)$  in Lemma A.7,*

$$\xi_1(\theta) = \bar{\theta}_{\mathcal{M},1} + o_p(1/\sqrt{n}),$$

where  $\bar{\theta}_{\mathcal{M},1} = \theta_{0,1} - (\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})^{-1} \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \bar{g}(\mathbf{D}, \theta_0)$  is the same as in Lemma A.2. Therefore,

$$\frac{n}{2}(\theta - \hat{\theta})^\top \Sigma(\theta)(\theta - \hat{\theta}) = \frac{n}{2}(\theta_1 - \bar{\theta}_{\mathcal{M},1})^\top \Sigma_{11}(\theta)(\theta_1 - \bar{\theta}_{\mathcal{M},1}) + T_{\mathcal{M}}(\theta) + o_p(1).$$

**Proof:** First we use the continuity of  $\mathbf{G}^\top(\theta)\mathbf{G}(\theta)$  and  $\mathbf{V}(\theta)$  in  $\theta$  from Assumption 10(iii), and replace  $\Sigma_{11}(\theta)$  and  $\Sigma_{12}(\theta)$  in the expression of  $\xi_1(\theta)$  by  $\Sigma_{11}$  and  $\Sigma_{12}$ , respectively. This is because we are considering  $\theta \in B_0(C\epsilon_n)$ , and this leads to (note that  $\hat{\theta}_2 = O_p(1/\sqrt{n})$ )

$$\xi_1(\theta) = \hat{\theta}_1 + (\Sigma_{11}^{-1}\Sigma_{12} + o_p(1))\hat{\theta}_2 = \hat{\theta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\hat{\theta}_2 + o_p(1/\sqrt{n}).$$

Next we use Assumption 10(v), and replace  $\hat{\theta}_1$  and  $\hat{\theta}_2$  with their first order approximations. For the  $\bar{\theta}$  in Assumption 10(v), suppose we decompose it into  $\bar{\theta} = (\bar{\theta}_1^\top, \bar{\theta}_2^\top)^\top$  according to a given model  $\mathcal{M}$ . Then by Assumption 10(v), we have  $\|\hat{\theta}_1 - \bar{\theta}_1\| = O_p(1/n)$  and  $\|\hat{\theta}_2 - \bar{\theta}_2\| = O_p(1/n)$ . Furthermore, through pure matrix algebra, we can derive that

$$\begin{aligned}\xi_1(\theta) &= \hat{\theta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\hat{\theta}_2 + o_p(1/\sqrt{n}) \\ &= \bar{\theta}_1 + \Sigma_{11}^{-1}\Sigma_{12}\bar{\theta}_2 + o_p(1/\sqrt{n}) + O_p(1/n) \\ &= (\mathbf{I}_{|\mathcal{M}|}, \Sigma_{11}^{-1}\Sigma_{12})\bar{\theta} + o_p(1/\sqrt{n}) \\ &= \bar{\theta}_{\mathcal{M},1} + o_p(1/\sqrt{n}),\end{aligned}$$

where in the last display,  $\bar{\theta}_{\mathcal{M},1}$  is defined in Lemma A.2. Therefore from Lemma A.7, for any  $\theta \in B_0(C\epsilon_n)$

$$\begin{aligned}& \frac{n}{2}(\theta - \hat{\theta})^\top \Sigma(\theta)(\theta - \hat{\theta}) \\ &= \frac{n}{2}(\theta_1 - \xi_1(\theta))^\top \Sigma_{11}(\theta)(\theta_1 - \xi_1(\theta)) + T_{\mathcal{M}}(\theta) \\ &= \frac{n}{2}(\theta_1 - \bar{\theta}_{\mathcal{M},1} + o_p(1/\sqrt{n}))^\top \Sigma_{11}(\theta)(\theta_1 - \bar{\theta}_{\mathcal{M},1} + o_p(1/\sqrt{n})) + T_{\mathcal{M}}(\theta) \\ &= \frac{n}{2}(\theta_1 - \bar{\theta}_{\mathcal{M},1})^\top \Sigma_{11}(\theta)(\theta_1 - \bar{\theta}_{\mathcal{M},1}) + T_{\mathcal{M}}(\theta) + o_p(1),\end{aligned}$$

which completes the proof. ■

**Lemma A.9.** Suppose Assumptions 1-10 hold. Let  $T_{\mathcal{M}}(\theta)$  be given in Lemma A.7. For any generic model  $\mathcal{M}$ , w.p.a.1,

(i) uniformly for all  $\theta \in B_0(C\epsilon_n) \cap \Theta(\mathcal{M})$  and all models  $\mathcal{M} \supseteq \mathcal{M}_0$ ,  $T_{\mathcal{M}}(\theta) - T_{\mathcal{M}}(\theta_0) = o_p(1)$ , given any fixed constant  $C > 0$ ;

(ii) for any  $\mathcal{M} \supseteq \mathcal{M}_0$ ,  $T_{\mathcal{M}_0}(\theta_0) - T_{\mathcal{M}}(\theta_0) = \log \frac{S_{\mathcal{M}}(\mathbf{D})}{S_{\mathcal{M}_0}(\mathbf{D})} + o_p(1)$ , where  $S_{\mathcal{M}}(\mathbf{D})$  is defined in Lemma A.2.

**Proof:** First, we can use Assumption 9 and the uniform boundedness of the eigenvalues of  $\mathbf{G}(\theta)^\top \mathbf{G}(\theta)$  and  $\mathbf{V}(\theta)$  for  $\theta \in \Theta$ , and express  $T_{\mathcal{M}}(\theta)$  as

$$T_{\mathcal{M}}(\theta)$$

$$\begin{aligned}
&= \frac{n}{2} \bar{\theta}_2^\top \mathbf{G}_{\mathcal{M}^c}(\theta)^\top \mathbf{J}_{\mathcal{M}}(\theta) \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}^c}(\theta) \bar{\theta}_2 + n \bar{\theta}_2^\top \mathbf{G}_{\mathcal{M}^c(\theta)}^\top \mathbf{J}_{\mathcal{M}}(\theta) \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}^c}(\theta) (\hat{\theta}_2 - \bar{\theta}_2) \\
&\quad + \frac{n}{2} (\hat{\theta}_2 - \bar{\theta}_2)^\top \mathbf{G}_{\mathcal{M}^c}(\theta)^\top \mathbf{J}_{\mathcal{M}}(\theta) \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}^c}(\theta) (\hat{\theta}_2 - \bar{\theta}_2) \\
&= \frac{n}{2} \bar{\theta}_2^\top \mathbf{G}_{\mathcal{M}^c}(\theta)^\top \mathbf{J}_{\mathcal{M}}(\theta) \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}^c}(\theta) \bar{\theta}_2 + O_p(1/\sqrt{n}) + O_p(1/n) \\
&= \frac{n}{2} \bar{\theta}_2^\top \mathbf{G}_{\mathcal{M}^c}(\theta)^\top \mathbf{J}_{\mathcal{M}}(\theta) \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}^c}(\theta) \bar{\theta}_2 + o_p(1), \tag{A.22}
\end{aligned}$$

where  $\bar{\theta}$  is the same as defined in Theorem 1 and  $\bar{\theta}_2$  is the subvector of  $\bar{\theta}$  with all those components not contained in model  $\mathcal{M}$ .

Furthermore, if  $\mathcal{M} \supseteq \mathcal{M}_0$ , then  $\theta_{0,2} = 0$ . It can be shown by straightforward matrix algebra that in this case,

$$\bar{\theta}_2 = (\mathbf{G}_{\mathcal{M}^c}^\top \mathbf{J}_{\mathcal{M}} \mathbf{V}^{-1} \mathbf{G}_{\mathcal{M}^c})^{-1} \mathbf{G}_{\mathcal{M}^c}^\top \mathbf{J}_{\mathcal{M}} \mathbf{V}^{-1} \bar{g}(\mathbf{D}, \theta_0). \tag{A.23}$$

Given the expression of  $T_{\mathcal{M}}(\theta)$  in (A.22) and  $\bar{\theta}_2$  in (A.23), it is clear that the dependence of  $T_{\mathcal{M}}(\theta)$  on  $\theta$  is only through the weighting matrix  $\mathbf{G}_{\mathcal{M}^c}(\theta)^\top \mathbf{J}_{\mathcal{M}}(\theta) \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}^c}(\theta)$  up to an error of  $o_p(1)$  that is uniform for all  $\mathcal{M}$ . Since Assumption 10(iii) has assumed the continuity of  $\mathbf{G}(\theta)$  and  $\mathbf{V}(\theta)$  with respect to  $\theta$ , it follows that for any fixed model  $\mathcal{M}$ ,  $\mathbf{G}_{\mathcal{M}^c}(\theta)^\top \mathbf{J}_{\mathcal{M}}(\theta) \mathbf{V}(\theta)^{-1} \mathbf{G}_{\mathcal{M}^c}(\theta)$  is also continuous in  $\theta$ . Moreover, due to the boundedness of  $p$  in Assumption 10(i), we have at most  $2^{\bar{p}}$  models  $\mathcal{M}$ , so the continuity is uniform in  $\mathcal{M}$ . Therefore, in the shrinking neighborhood  $\theta \in B_0(C\epsilon_n)$  where  $\epsilon_n \rightarrow 0$ , we have that uniformly over all models  $\mathcal{M} \supseteq \mathcal{M}_0$ ,  $T_{\mathcal{M}}(\theta) - T_{\mathcal{M}}(\theta_0) = o_p(1)$ , which has proved (i).

By (A.22) and (A.23), for any  $\mathcal{M} \supseteq \mathcal{M}_0$ , we have  $T_{\mathcal{M}}(\theta_0) = \frac{n}{2} \bar{g}(\mathbf{D}, \theta_0)^\top \mathbf{K}(\mathcal{M}) \bar{g}(\mathbf{D}, \theta_0) + o_p(1)$ , where

$$\mathbf{K}(\mathcal{M}) = \mathbf{V}^{-1} \mathbf{J}_{\mathcal{M}}^\top \mathbf{G}_{\mathcal{M}^c} (\mathbf{G}_{\mathcal{M}^c}^\top \mathbf{J}_{\mathcal{M}} \mathbf{V}^{-1} \mathbf{G}_{\mathcal{M}^c})^{-1} \mathbf{G}_{\mathcal{M}^c}^\top \mathbf{J}_{\mathcal{M}} \mathbf{V}^{-1}.$$

By pure matrix algebra, it can be shown that

$$\begin{aligned}
&\mathbf{K}(\mathcal{M}_0) - \mathbf{K}(\mathcal{M}) \\
&= \mathbf{V}^{-1} \mathbf{G}_{\mathcal{M}} (\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}^{-1} \mathbf{G}_{\mathcal{M}})^{-1} \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{G}_{\mathcal{M}_0} (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1} \mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}^{-1}
\end{aligned}$$

Because  $\mathbf{V}_n$  is a consistent estimator for  $\mathbf{V}$  and their eigenvalues are bounded, we can replace  $\mathbf{V}$  in the display above by  $\mathbf{V}_n$ , which will incur an error of  $o_p(1)$ . Then it follows that

$$\begin{aligned}
&T_{\mathcal{M}_0}(\theta_0) - T_{\mathcal{M}}(\theta_0) \\
&= \frac{n}{2} \bar{g}(\mathbf{D}, \theta_0)^\top \left[ \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}} (\mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}})^{-1} \mathbf{G}_{\mathcal{M}}^\top \mathbf{V}_n^{-1} \right.
\end{aligned}$$

$$- \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0} (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1} \mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \Big] \bar{g}(\mathbf{D}, \theta_0) + o_p(1).$$

We compare this with (A.11) and obtain that  $T_{\mathcal{M}_0}(\theta_0) - T_{\mathcal{M}}(\theta_0) = \log \frac{S_{\mathcal{M}}(\mathbf{D})}{S_{\mathcal{M}_0}(\mathbf{D})} + o_p(1)$ . (ii) is proved.  $\blacksquare$

**Lemma A.10.** *Under Assumptions 1-11, uniformly for any generic model  $\mathcal{M}$ ,*

$$\int_{\Theta(\mathcal{M})} \left| p(\hat{\theta}|\theta_1) - \phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) \right| \pi(\theta_1|\mathcal{M}) d\theta_1 = O_p\left(\tau_n n^{\frac{p-|\mathcal{M}|}{2}}\right), \quad (\text{A.24})$$

where

$$\begin{aligned} & \phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) \\ &= \left(\frac{2\pi}{n}\right)^{-\frac{p}{2}} \det(\Sigma(\theta))^{1/2} \exp \left\{ -\frac{n}{2} (\hat{\theta}_1^\top - \theta_1^\top, \hat{\theta}_2^\top) \Sigma(\theta) \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 \end{pmatrix} \right\}. \end{aligned} \quad (\text{A.25})$$

**Proof:** Let  $c(x) = 1/(1+x^{p+1})$ . First one can do a variable transformation from  $\hat{\theta}$  to  $Z = \sqrt{n} \mathbf{F}(\theta)(\hat{\theta} - \theta)$ , where  $\theta = (\theta_1^\top, 0)^\top$  and  $\mathbf{F}(\theta)^\top \mathbf{F}(\theta) = \mathbf{G}(\theta)^\top \mathbf{V}(\theta)^{-1} \mathbf{G}(\theta) = \Sigma(\theta)$ . The densities have the relation  $p(\hat{\theta}|\theta_1) = n^{p/2} \det(\mathbf{F}(\theta)) p_Z(z|\theta_1)$  and  $\phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) = n^{p/2} \det(\mathbf{F}(\theta)) \phi(z; 0, \mathbf{I}_p)$ . Note that this transformation is in  $\mathbb{R}^p$  and does not directly involve the integration with respect to  $\theta_1$ . Then using Assumption 10, the left-hand side of (A.24) can be bounded by

$$\begin{aligned} & \int_{\Theta(\mathcal{M})} \left| p(\hat{\theta}|\theta_1) - \phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) \right| \pi(\theta_1|\mathcal{M}) d\theta_1 \\ &= n^{p/2} \sup_{\theta \in \Theta} \det(\mathbf{F}(\theta)) \int_{\Theta(\mathcal{M})} \left| p_Z(z|\theta_1) - \phi(z; 0, \mathbf{I}_p) \right| \pi(\theta_1|\mathcal{M}) d\theta_1 \\ &\leq n^{p/2} \sup_{\theta \in \Theta} \left( \frac{\bar{\lambda}(\mathbf{G}(\theta)^\top \mathbf{G}(\theta))}{\underline{\lambda}(\mathbf{V}(\theta))} \right)^{p/2} \cdot \tau_n \int_{\Theta(\mathcal{M})} c(\|z\|) \pi(\theta_1|\mathcal{M}) d\theta_1 \\ &\leq n^{p/2} \sup_{\theta \in \Theta} \left( \frac{\bar{\lambda}(\mathbf{G}(\theta)^\top \mathbf{G}(\theta))}{\underline{\lambda}(\mathbf{V}(\theta))} \right)^{p/2} \cdot \tau_n c_\pi \int_{\Theta(\mathcal{M})} c(\|z\|) d\theta_1, \end{aligned} \quad (\text{A.26})$$

where in the last inequality we have used Assumption 7(i) that  $\pi(\theta_1|\mathcal{M}) \leq c_\pi$  uniformly for all  $\mathcal{M}$ . We now transform  $\theta_1$  into  $\beta = \sqrt{n}(\hat{\theta}_1 - \theta_1)$  in the foregoing integration about the function  $c(\|z\|)$ . We have

$$\|z\|^2 = n(\hat{\theta} - \theta)^\top \Sigma(\theta)(\hat{\theta} - \theta) \geq \underline{\lambda}(\Sigma(\theta)) \|\sqrt{n}(\hat{\theta} - \theta)\|^2,$$

and

$$\|\sqrt{n}(\hat{\theta} - \theta)\|^2 = \|\sqrt{n}(\hat{\theta}_1 - \theta_1)\|^2 + n\|\hat{\theta}_2\|^2 = \|\beta\|^2 + n\|\hat{\theta}_2\|^2 \geq \|\beta\|^2.$$

Furthermore, the eigenvalue satisfies

$$\inf_{\theta \in \Theta} \underline{\lambda}(\Sigma(\theta)) \geq \inf_{\theta \in \Theta} \underline{\lambda}(\mathbf{G}(\theta)^\top \mathbf{G}(\theta)) \bar{\lambda}(\mathbf{V}(\theta))^{-1},$$

which is lower bounded by constant according to Assumption 10(iii). Therefore, along with the nonincreasing property of the function  $c(\cdot)$ , we have

$$\begin{aligned} \int_{\Theta(\mathcal{M})} c(\|z\|) d\theta_1 &\leq n^{-\frac{|\mathcal{M}|}{2}} \int_{\mathbb{R}^{|\mathcal{M}|}} c\left(\inf_{\theta \in \Theta} \sqrt{\frac{\underline{\lambda}(\mathbf{G}(\theta)^\top \mathbf{G}(\theta))}{\bar{\lambda}(\mathbf{V}(\theta))}} \|\beta\|\right) d\beta \\ &\leq n^{-\frac{|\mathcal{M}|}{2}} \left( \inf_{\theta \in \Theta} \frac{\underline{\lambda}(\mathbf{G}(\theta)^\top \mathbf{G}(\theta))}{\bar{\lambda}(\mathbf{V}(\theta))} \right)^{-\frac{|\mathcal{M}|}{2}} \int_0^\infty \frac{x^{|\mathcal{M}|-1}}{1+x^{p+1}} dx = O(n^{-\frac{|\mathcal{M}|}{2}}) \end{aligned} \quad (\text{A.27})$$

The conclusion follows from (A.26), (A.27), and the boundedness of the eigenvalues of  $\Sigma(\theta)$ . ■

### Proof of Theorem 5 (i):

We derive an order expression for  $\text{BF}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] = \frac{\int_{\Theta(\mathcal{M})} p(\hat{\theta}|\theta_1) \pi(\theta_1|\mathcal{M}) d\theta_1}{\int_{\Theta(\mathcal{M}_0)} p(\hat{\theta}|\theta_1) \pi(\theta_1|\mathcal{M}_0) d\theta_1}$  where  $\mathcal{M} \supseteq \mathcal{M}_0$  (note that the two  $\theta_1$  in the numerator and the denominator have different dimensions). Hereafter, all  $o_p$  and  $O_p$  hold uniformly over all the models with  $\mathcal{M} \supseteq \mathcal{M}_0$ . First of all, based on Lemma A.10, we have

$$\begin{aligned} &\int_{\Theta(\mathcal{M})} p(\hat{\theta}|\theta_1) \pi(\theta_1|\mathcal{M}) d\theta_1 \\ &\leq \int_{\Theta(\mathcal{M})} \left| p(\hat{\theta}_1|\theta_1) - \phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) \right| \pi(\theta_1|\mathcal{M}) d\theta_1 + \int_{\Theta(\mathcal{M})} \phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) \pi(\theta_1|\mathcal{M}) d\theta_1 \\ &\leq O_p\left(\tau_n n^{\frac{p-|\mathcal{M}|}{2}}\right) + \int_{\Theta(\mathcal{M})} \phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) \pi(\theta_1|\mathcal{M}) d\theta_1 \end{aligned} \quad (\text{A.28})$$

We claim that the second term in (A.28) satisfies

$$\int_{\Theta(\mathcal{M})} \phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) \pi(\theta_1|\mathcal{M}) d\theta_1 - \int_{\Theta(\mathcal{M})} \phi(\hat{\theta}; \theta_1, \Sigma^{-1}/n) \pi(\theta_1|\mathcal{M}) d\theta_1 = o_p\left(n^{\frac{p-|\mathcal{M}|}{2}}\right), \quad (\text{A.29})$$

where we have replaced  $\Sigma(\theta)$  with  $\Sigma$ , i.e. the matrix  $\Sigma(\theta)$  evaluated at  $\theta = \theta_0$ .

To show (A.29), we first observe that both integrals in the display can be made to order  $o_p(n^{\frac{p-|\mathcal{M}|}{2}})$  outside the neighborhood  $B_0(C\epsilon_n)$  for some constant  $C > 0$ . This is because by the decomposition in Lemma A.7,

$$\begin{aligned} &\int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} \phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) \pi(\theta_1|\mathcal{M}) d\theta_1 \\ &= \int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} \left(\frac{2\pi}{n}\right)^{-p/2} \det(\Sigma(\theta))^{1/2} e^{-\frac{n}{2}(\theta - \hat{\theta})^\top \Sigma(\theta)(\theta - \hat{\theta})} \pi(\theta_1|\mathcal{M}) d\theta_1 \end{aligned}$$

$$\begin{aligned}
&\leq c_\pi \left(\frac{2\pi}{n}\right)^{-\frac{p}{2}} \sup_{\theta \in \Theta} \det(\Sigma(\theta))^{1/2} \int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} \exp \left\{ -\frac{n \inf_{\theta \in \Theta} \underline{\lambda}(\Sigma(\theta))}{2} (\theta - \hat{\theta})^\top (\theta - \hat{\theta}) \right\} d\theta_1 \\
&\leq c_\pi \left(\frac{2\pi}{n}\right)^{-\frac{p}{2}} \sup_{\theta \in \Theta} \det(\Sigma(\theta))^{1/2} \int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} \exp \left\{ -\frac{n \inf_{\theta \in \Theta} \underline{\lambda}(\Sigma(\theta))}{2} (\theta_1 - \hat{\theta}_1)^\top (\theta_1 - \hat{\theta}_1) \right\} d\theta_1 \\
&\leq c_\pi \left(\frac{2\pi}{n}\right)^{-\frac{p-|\mathcal{M}|}{2}} \frac{\sup_{\theta \in \Theta} \det(\Sigma(\theta))^{1/2}}{\inf_{\theta \in \Theta} \underline{\lambda}(\Sigma(\theta))^{|\mathcal{M}|/2}} \int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} \phi\left(\theta_1; \hat{\theta}_1, \frac{1}{n \inf_{\theta \in \Theta} \underline{\lambda}(\Sigma(\theta))} \mathbf{I}_{|\mathcal{M}|}\right) d\theta_1
\end{aligned}$$

where we have used Assumption 7(i) that the prior is bounded, and the fact that  $T_{\mathcal{M}}(\theta) \geq 0$ . All the determinants and the eigenvalues here are bounded from below and above, by Assumption 10. By choosing  $C$  sufficiently large, we can make the integral in the last display arbitrarily small, due to the Gaussian concentration inequality. Hence as we choose  $C$  arbitrarily large,

$$\int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} \phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) \pi(\theta_1 | \mathcal{M}) d\theta_1 = o_p\left(n^{\frac{p-|\mathcal{M}|}{2}}\right). \quad (\text{A.30})$$

Similarly we have

$$\int_{\Theta(\mathcal{M}) \setminus B_0(C\epsilon_n)} \phi(\hat{\theta}; \theta_1, \Sigma^{-1}/n) \pi(\theta_1 | \mathcal{M}) d\theta_1 = o_p\left(n^{\frac{p-|\mathcal{M}|}{2}}\right). \quad (\text{A.31})$$

Therefore it is sufficient to show (A.29) on  $\Theta(\mathcal{M}) \cap B_0(C\epsilon_n)$ . We will use the continuity of  $\Sigma(\theta)$  with respect to  $\theta$  again, in the sense that  $\det(\Sigma(\theta))/\det(\Sigma) = 1 + o_p(1)$  and also  $\|\Sigma_{11}(\theta) - \Sigma_{11}\| = o_p(1)$  for  $\theta \in B_0(C\epsilon_n)$ . It follows from Lemma A.8, Lemma A.9 (i) and the Gaussian concentration inequality that

$$\begin{aligned}
&\int_{\Theta(\mathcal{M}) \cap B_0(C\epsilon_n)} \phi(\hat{\theta}; \theta_1, \Sigma^{-1}(\theta)/n) \pi(\theta_1 | \mathcal{M}) d\theta_1 \\
&= \int_{\Theta(\mathcal{M}) \cap B_0(C\epsilon_n)} \left(\frac{2\pi}{n}\right)^{-p/2} \det(\Sigma(\theta))^{1/2} e^{-\frac{n}{2}(\theta_1 - \bar{\theta}_{\mathcal{M},1})^\top \Sigma_{11}(\theta)(\theta_1 - \bar{\theta}_{\mathcal{M},1}) - T_{\mathcal{M}}(\theta) + o_p(1)} \pi(\theta_1 | \mathcal{M}) d\theta_1 \\
&= \int_{\Theta(\mathcal{M}) \cap B_0(C\epsilon_n)} \left(\frac{2\pi}{n}\right)^{-p/2} \det(\Sigma)^{1/2} (1 + o_p(1)) \cdot \\
&\quad \exp \left\{ -\frac{n}{2}(\theta_1 - \bar{\theta}_{\mathcal{M},1})^\top \Sigma_{11}(\theta)(\theta_1 - \bar{\theta}_{\mathcal{M},1}) + o_p\left[\frac{n}{2}\|\theta_1 - \bar{\theta}_{\mathcal{M},1}\|^2\right] \right. \\
&\quad \left. - T_{\mathcal{M}}(\theta) + o_p(1) \right\} \pi(\theta_1 | \mathcal{M}) d\theta_1 \\
&= (1 + o_p(1)) \int_{\Theta(\mathcal{M}) \cap B_0(C\epsilon_n)} \left(\frac{2\pi}{n}\right)^{-p/2} \det(\Sigma)^{1/2} \cdot \\
&\quad \exp \left\{ -\frac{n}{2}(\theta_1 - \bar{\theta}_{\mathcal{M},1})^\top \Sigma_{11}(\theta)(\theta_1 - \bar{\theta}_{\mathcal{M},1}) - T_{\mathcal{M}}(\theta) \right\} \pi(\theta_1 | \mathcal{M}) d\theta_1 \\
&= (1 + o_p(1)) \int_{\Theta(\mathcal{M}) \cap B_0(C\epsilon_n)} \phi(\hat{\theta}; \theta_1, \Sigma^{-1}/n) \pi(\theta_1 | \mathcal{M}) d\theta_1 \\
&= (1 + o_p(1)) e^{-T_{\mathcal{M}}(\theta_0)} \left(\frac{2\pi}{n}\right)^{-\frac{p-|\mathcal{M}|}{2}} \left(\frac{\det(\Sigma)}{\det(\Sigma_{11})}\right)^{1/2} \pi(\theta_{0,1} | \mathcal{M}). \quad (\text{A.32})
\end{aligned}$$

Therefore, (A.29) follows immediately from (A.30) - (A.32), because  $T_{\mathcal{M}}(\theta_0) = O_p(1)$ , and Assumption 7 (iii) now implies a constant lower bound for  $\pi(\theta_{0,1}|\mathcal{M})$ . We can further combine (A.28) and (A.29) and conclude that

$$\int_{\Theta(\mathcal{M})} p(\hat{\theta}|\theta_1)\pi(\theta_1|\mathcal{M})d\theta_1 = \int_{\Theta(\mathcal{M})} \phi(\hat{\theta}; \theta_1, \Sigma^{-1}/n)\pi(\theta_1|\mathcal{M})d\theta_1 + o_p\left(n^{\frac{p-|\mathcal{M}|}{2}}\right).$$

Given this result, the Bayes factor based on  $p(\hat{\theta}|\theta_1)$  can be directly translated into the Bayes factor based on  $\phi(\hat{\theta}; \theta_1, \Sigma^{-1}/n)$ . It follows that

$$\begin{aligned} & \text{BF}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] \\ &= \frac{\int_{\Theta(\mathcal{M})} p(\hat{\theta}|\theta_1)\pi(\theta_1|\mathcal{M})d\theta_1}{\int_{\Theta(\mathcal{M}_0)} p(\hat{\theta}|\theta_1)\pi(\theta_1|\mathcal{M}_0)d\theta_1} \\ &= \frac{\int_{\Theta(\mathcal{M})} \phi(\hat{\theta}; \theta_1, \Sigma^{-1}/n)\pi(\theta_1|\mathcal{M})d\theta_1 + o_p\left(n^{\frac{p-|\mathcal{M}|}{2}}\right)}{\int_{\Theta(\mathcal{M}_0)} \phi(\hat{\theta}; \theta_1, \Sigma^{-1}/n)\pi(\theta_1|\mathcal{M}_0)d\theta_1 + o_p\left(n^{\frac{p-|\mathcal{M}_0|}{2}}\right)} \\ &\stackrel{\text{(A.30)-(A.32)}}{=} \frac{(1 + o_p(1))e^{-T_{\mathcal{M}}(\theta_0)}\left(\frac{2\pi}{n}\right)^{-\frac{p-|\mathcal{M}|}{2}}\det(\mathbf{G}_{\mathcal{M}}^{\top}\mathbf{V}^{-1}\mathbf{G}_{\mathcal{M}})^{-1/2}\pi(\theta_{0,1}|\mathcal{M}) + o_p\left(n^{\frac{p-|\mathcal{M}|}{2}}\right)}{(1 + o_p(1))e^{-T_{\mathcal{M}_0}(\theta_0)}\left(\frac{2\pi}{n}\right)^{-\frac{p-|\mathcal{M}_0|}{2}}\det(\mathbf{G}_{\mathcal{M}_0}^{\top}\mathbf{V}^{-1}\mathbf{G}_{\mathcal{M}_0})^{-1/2}\pi(\theta_{0,1}|\mathcal{M}_0) + o_p\left(n^{\frac{p-|\mathcal{M}_0|}{2}}\right)} \\ &\stackrel{\text{Lemma A.9 (ii)}}{=} (1 + o_p(1))\left(\frac{2\pi}{n}\right)^{-\frac{|\mathcal{M}|-|\mathcal{M}_0|}{2}}\frac{S_{\mathcal{M}}(\mathbf{D})}{S_{\mathcal{M}_0}(\mathbf{D})} \cdot \frac{[\det(\mathbf{G}_{\mathcal{M}}^{\top}\mathbf{V}_n^{-1}\mathbf{G}_{\mathcal{M}})]^{-1/2}\pi(\theta_0|\mathcal{M})}{[\det(\mathbf{G}_{\mathcal{M}_0}^{\top}\mathbf{V}_n^{-1}\mathbf{G}_{\mathcal{M}_0})]^{-1/2}\pi(\theta_0|\mathcal{M}_0)}. \end{aligned}$$

The last display is exactly the expression of  $\text{BF}_q[\mathcal{M} : \mathcal{M}_0]$  in Lemma A.5. Hence  $\frac{\text{BF}_q[\mathcal{M}:\mathcal{M}_0]}{\text{BF}_{\hat{\theta}}[\mathcal{M}:\mathcal{M}_0]} \rightarrow 1$  has been proved. Since now  $p$  is bounded above by  $\bar{p}$  in Assumption 10(i), one can see that the order of  $\text{BF}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0]$  is equal to  $n^{-\frac{|\mathcal{M}|-k_0}{2}}$ , which completes the proof.  $\blacksquare$

### Proof of Theorem 5 (ii):

The first inequality directly follows from Lemma A.6 and Assumption 10(ii). For the second one, we have that for  $z = \sqrt{n}\mathbf{F}(\theta)(\hat{\theta} - \theta)$  with any  $\theta \in \Theta(\mathcal{M})$  and  $\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset$ ,

$$\begin{aligned} \|z\|^2 &= n(\hat{\theta} - \theta)^{\top}\Sigma(\theta)(\hat{\theta} - \theta) \geq \inf_{\theta \in \Theta} \underline{\lambda}(\Sigma(\theta))n\|\hat{\theta} - \theta\|^2 \\ &\geq n \inf_{\theta \in \Theta} \underline{\lambda}(\Sigma(\theta))[\|\theta - \hat{\theta}_1\|^2 + \|\hat{\theta}_2\|^2] \end{aligned}$$

Since  $\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset$ , in large sample,  $\hat{\theta}_2 \rightarrow \theta_{0,2} \neq 0$ . Furthermore, by Assumption 10(ii),

$$\|\hat{\theta}_2\|^2 = \|\theta_{0,2}\|^2 + o_p(1) \geq \frac{1}{2}\|\theta_{0,2}\|^2 \geq \frac{1}{2}\underline{\theta}^2.$$

Therefore, if we let  $C_1 = \inf_{\theta \in \Theta} \underline{\lambda}(\Sigma(\theta))$ , then as  $n \rightarrow \infty$ , w.p.a.1,

$$\|z\|^2 \geq C_1 n[\|\theta - \hat{\theta}_1\|^2 + \|\hat{\theta}_2\|^2] \geq C_1 n\|\theta - \hat{\theta}_1\|^2 + \frac{C_1 \underline{\theta}^2 n}{2}.$$



Thus for some constant  $C > 0$ , we can bound the difference between  $p(\hat{\theta}|\theta_1)$  and the normal limit, by

$$\begin{aligned}
& \int_{\Theta(\mathcal{M})} \left| p(\hat{\theta}|\theta_1) - \phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) \right| \pi(\theta_1|\mathcal{M}) d\theta_1 \\
& \leq n^{p/2} \sup_{\theta \in \Theta} (\det(\Sigma(\theta)))^{1/2} \cdot \tau_n c_\pi \int_{\Theta(\mathcal{M})} \frac{1}{1 + (\|z\|^2)^{\frac{p+1}{2}}} d\theta_1 \\
& \leq C \tau_n n^{p/2} \int_{\Theta(\mathcal{M})} \frac{1}{[C_1 n \|\theta_1 - \hat{\theta}_1\|^2 + \frac{C_1 \theta^2 n}{2}]^{\frac{p+1}{2}}} d\theta_1 \\
& = C \tau_n n^{p/2} \int_{\Theta(\mathcal{M})} \frac{(C_1 \theta^2 n/2)^{-\frac{p+1}{2}}}{[\frac{\|\theta_1 - \hat{\theta}_1\|^2}{\theta^2/2} + 1]^{\frac{p+1}{2}}} d\theta_1 \\
& \leq C \tau_n n^{-\frac{1}{2}} \int_{\Theta(\mathcal{M})} \frac{1}{(1 + \|u\|^2)^{\frac{p+1}{2}}} du \\
& \leq C \tau_n n^{-\frac{1}{2}} \int_0^\infty \frac{x^{|\mathcal{M}|-1}}{(1+x^2)^{\frac{p+1}{2}}} dx \\
& \leq C \tau_n n^{-\frac{1}{2}}, \tag{A.33}
\end{aligned}$$

where  $C$  has absorbed all the constant terms. Using (A.33), we can bound the marginal probability  $\int_{\Theta(\mathcal{M})} p(\hat{\theta}|\theta_1) \pi(\theta_1|\mathcal{M}) d\theta_1$  as

$$\begin{aligned}
& \int_{\Theta(\mathcal{M})} p(\hat{\theta}|\theta_1) \pi(\theta_1|\mathcal{M}) d\theta_1 \\
& \leq \int_{\Theta(\mathcal{M})} \left| p(\hat{\theta}|\theta_1) - \phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) \right| \pi(\theta_1|\mathcal{M}) d\theta_1 + \int_{\Theta(\mathcal{M})} \phi(\hat{\theta}; \theta_1, \Sigma(\theta)^{-1}/n) \pi(\theta_1|\mathcal{M}) d\theta_1 \\
& \leq C \tau_n n^{-\frac{1}{2}} + c_\pi \int_{\Theta(\mathcal{M})} \left( \frac{2\pi}{n} \right)^{-\frac{p}{2}} \det(\Sigma(\theta))^{1/2} e^{-\frac{C_1 n}{2} (\theta_1 - \hat{\theta}_1)^\top (\theta_1 - \hat{\theta}_1) - \frac{C_1 n \theta^2}{4}} d\theta_1 \\
& \leq C \tau_n n^{-\frac{1}{2}} + c_\pi \det(\Sigma(\theta))^{1/2} C_1^{-|\mathcal{M}|/2} e^{-C_1 n \theta^2/4} \left( \frac{2\pi}{n} \right)^{-\frac{p-|\mathcal{M}|}{2}} \\
& \leq C_1 \tau_n n^{-\frac{1}{2}} + e^{-C_2 n \theta^2}
\end{aligned}$$

for some redefined constants  $C_1, C_2 > 0$ . Therefore, the Bayes factor can be bounded w.p.a.1 by

$$\begin{aligned}
& \text{BF}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] \\
& = \frac{\int_{\Theta(\mathcal{M})} p(\hat{\theta}|\theta_1) \pi(\theta_1|\mathcal{M}) d\theta_1}{\int_{\Theta(\mathcal{M}_0)} p(\hat{\theta}|\theta_1) \pi(\theta_1|\mathcal{M}_0) d\theta_1} \\
& = \frac{C_1 \tau_n n^{-\frac{1}{2}} + e^{-C_2 n \theta^2}}{(1 + o_p(1)) e^{-T_{\mathcal{M}_0}(\theta_0)} \left( \frac{2\pi}{n} \right)^{-\frac{p-|\mathcal{M}_0|}{2}} \det(\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1/2} \pi(\theta_{0,1}|\mathcal{M}_0) + o_p(n^{\frac{p-|\mathcal{M}_0|}{2}})} \\
& \leq \tau_n n^{\frac{k_0-p-1}{2}} + e^{-C n \theta^2}
\end{aligned}$$

for some redefined constant  $C > 0$  and the first constant can be absorbed into  $\tau_n$ . This completes the proof of Theorem 5 (ii).  $\blacksquare$

**Proof of Theorem 6 (i):**

If  $\mathcal{M}_0 \neq \mathcal{M}_{\text{full}}$ , then there exists at least one model  $\mathcal{M}$  such that  $\mathcal{M} \supset \mathcal{M}_0$ . Also the total number of models is now bounded by  $2^{\bar{p}}$ . Hence under the same prior  $\pi(\theta, \mathcal{M})$ , Theorem 5 (i) and Assumption 10(iv) imply that  $\sum_{\mathcal{M}:\mathcal{M} \supset \mathcal{M}_0} \text{PO}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] = O_p(n^{-1/2})$ ,  $\sum_{\mathcal{M}:\mathcal{M} \supset \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] = O_p(n^{-1/2})$ , and

$$\sum_{\mathcal{M}:\mathcal{M} \supset \mathcal{M}_0} \text{PO}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] = (1 + o_p(1)) \sum_{\mathcal{M}:\mathcal{M} \supset \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0].$$

On the other hand, Theorem 5 (ii) implies that  $\sum_{\mathcal{M}:\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] = \exp(-Cn\underline{\theta}^2)$  for some constant  $C > 0$ , and also

$$\begin{aligned} & \sum_{\mathcal{M}:\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset} \text{PO}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] \\ & \leq 2^p \exp(-Cn\underline{\theta}^2) \vee \tau_n n^{\frac{k_0 - p - 1}{2}} \\ & \leq \exp(-Cn\underline{\theta}^2) \vee \tau_n n^{\frac{k_0 - p - 1}{2}} \end{aligned}$$

with adjusted  $C$  and  $\tau_n = o_p(1)$ . Therefore, it is clear that

$$\begin{aligned} \sum_{\mathcal{M}:\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] &= o_p\left(\sum_{\mathcal{M}:\mathcal{M} \supset \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0]\right) \\ \sum_{\mathcal{M}:\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset} \text{PO}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] &= o_p\left(\sum_{\mathcal{M}:\mathcal{M} \supset \mathcal{M}_0} \text{PO}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0]\right) \end{aligned}$$

Hence it is clear that the posterior consistency follows, with  $q(\mathcal{M}_0 | \mathbf{D}) = 1 + o_p(1)$  and  $p(\mathcal{M}_0 | \hat{\theta}) = 1 + o_p(1)$ . Moreover,

$$\begin{aligned} q(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0 | \mathbf{D}) &= \left( \sum_{\mathcal{M}:\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] + \sum_{\mathcal{M}:\mathcal{M} \supset \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] \right) q(\mathcal{M}_0 | \mathbf{D}) \\ &= (1 + o_p(1)) \sum_{\mathcal{M}:\mathcal{M} \supset \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] \asymp n^{-1/2} \\ p(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0 | \hat{\theta}) &= \left( \sum_{\mathcal{M}:\mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset} \text{PO}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] + \sum_{\mathcal{M}:\mathcal{M} \supset \mathcal{M}_0} \text{PO}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] \right) p(\mathcal{M}_0 | \hat{\theta}) \\ &= (1 + o_p(1)) \sum_{\mathcal{M}:\mathcal{M} \supset \mathcal{M}_0} \text{PO}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] \asymp n^{-1/2} \\ \frac{q(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0 | \mathbf{D})}{p(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0 | \hat{\theta})} &= \frac{(1 + o_p(1)) \sum_{\mathcal{M}:\mathcal{M} \supset \mathcal{M}_0} \text{PO}_q[\mathcal{M} : \mathcal{M}_0]}{(1 + o_p(1)) \sum_{\mathcal{M}:\mathcal{M} \supset \mathcal{M}_0} \text{PO}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0]} \rightarrow 1 \end{aligned}$$

as  $n \rightarrow \infty$ , w.p.a.1.

When  $\mathcal{M}_0 = \mathcal{M}_{\text{full}}$ , there is no model  $\mathcal{M}$  with  $\mathcal{M} \supset \mathcal{M}_0$ . Hence

$$\begin{aligned} q(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0 | \mathbf{D}) &= \sum_{\mathcal{M} : \mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset} \text{PO}_q[\mathcal{M} : \mathcal{M}_0] \cdot q(\mathcal{M}_0 | \mathbf{D}) \leq \exp(-Cn\underline{\theta}^2); \\ p(\mathcal{M} : \mathcal{M} \neq \mathcal{M}_0 | \hat{\theta}) &= \sum_{\mathcal{M} : \mathcal{M}_0 \setminus \mathcal{M} \neq \emptyset} \text{PO}_{\hat{\theta}}[\mathcal{M} : \mathcal{M}_0] \cdot p(\mathcal{M}_0 | \hat{\theta}) \leq \exp(-Cn\underline{\theta}^2) \vee \tau_n n^{\frac{k_0 - p - 1}{2}}. \end{aligned}$$

■

### Proof of Theorem 6 (ii):

Because of the model selection consistency in Theorem 6 (i) for  $p(\theta | \hat{\theta})$  and the normal approximation from (A.29)-(A.32), one can show that

$$\sup_{A \subseteq \Theta} \left| \int_A p(\theta | \hat{\theta}) d\theta - \int_A \phi(\theta; \hat{\theta}_{\mathcal{M}_0, 1}, (\mathbf{G}_{\mathcal{M}_0}^\top \mathbf{V}_n^{-1} \mathbf{G}_{\mathcal{M}_0})^{-1} / n) d\theta \right| \rightarrow 0.$$

The proof proceeds in the same manner as the proof of Theorem 1(ii), hence we omit it here. Based on Assumption 1-11, the result follows immediately by combining the conclusions of Corollary 1 and the display above.

■

## References

- P. Alquier and G. Biau. Sparse single-index model. *Journal of Machine Learning Research*, 14:243–280, 2013.
- D. W. K. Andrews. Consistent moment selection procedures for generalized method of moments estimation. *Econometrica*, 67:543–564, 1999.
- D. W. K. Andrews and B. Lu. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *Journal of Econometrics*, 101:123–164, 2001.
- A. Belloni and V. Chernozhukov. On the computational complexity of MCMC-based estimators in large samples. *The Annals of Statistics*, 37:2011–2055, 2009.
- A. Belloni, V. Chernozhukov, and I. Fernández-Val. Conditional quantile processes based on series or many regressors. arXiv: 1105.6154, 2011.
- R. N. Bhattacharya and R. Ranga Rao. *Normal Approximation and Asymptotic Expansions*. Wiley, New York 1976. Reprinted by Robert E. Krieger, Melbourne, Florida, 1986.
- P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer, Berlin, 2011.
- M. Caner and H. H. Zhang. Adaptive elastic net GMM estimator. *Journal of Business and Economics Statistics*, 32:30–47, 2013.
- M. Caner, X. Han, and Y. Lee. Adaptive elastic net GMM estimator with many invalid moment conditions: A simultaneous model and moment selection. Manuscript. 2013.
- J. Chen and Z. Chen. Extended Bayesian information criterion for model selection with large model space. *Biometrika*, 94:759–771, 2008.
- K. Chen, W. Jiang, and M. Tanner. A note on some algorithms for the Gibbs posterior. *Statistics and Probability Letters*, 80:1234–1241, 2010.
- X. Chen and Z. Liao. Select the valid and relevant moments: An information-based LASSO for GMM with many moments. Manuscript. 2013.

- V. Chernozhukov and C. Hansen. An IV model of quantile treatment effects. *Econometrica*, 73:245–261, 2005.
- V. Chernozhukov and C. Hansen. Instrumental quantile regression inference for structural and treatment effects model. *Journal of Econometrics*, 132:491–525, 2006.
- V. Chernozhukov and H. Hong. An MCMC approach to classical estimation. *Journal of Econometrics*, 115:293–346, 2003.
- H. Chipman, E. I. George, and R. E. McCulloch. The practical implementation of Bayesian model selection. *Model Selection, IMS Lecture Notes - Monograph Series*, 38:65–116, 2001.
- H. Cho and A. Qu. Model selection for correlated data with diverging number of parameters. *Statistica Sinica*, 23:901–927, 2013.
- P. Dellaportas, J. J. Forster, and I. Ntzoufras. On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12:27–36, 2002.
- M. Drton and M. D. Perlman. Model selection for Gaussian concentration graphs. *Biometrika*, 91:591–602, 2004.
- J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.
- K. T. Fang and R. Mukerjee. Empirical-type likelihoods allowing posterior credible sets with frequentist validity: Higher-order asymptotics. *Biometrika*, 93:723–733, 2006.
- J. P. Florens and A. Simoni. Gaussian processes and Bayesian moment estimation. Manuscript. 2012.
- A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman and Hall/CRC, 2013.
- E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.
- P. Green. Reversible jump Markov chain Monte Carlo. *Biometrika*, 82:711–732, 1995.
- B. Guedj and P. Alquier. PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, 7:264–291, 2013.

- L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054, 1982.
- L. P. Hansen, J. Heaton, and A. Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business and Economic Statistics*, 14:262–280, 1996.
- H. Hong and B. Preston. Bayesian averaging, prediction and nonnested model selection. *Journal of Econometrics*, 167:358–369, 2012.
- P. Huber. Robust regressions: asymptotics, conjectures, and Monte Carlo. *The Annals of Statistics*, 1:799–821, 1973.
- H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 33:730–773, 2005.
- W. Jiang. Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. *The Annals of Statistics*, 35:1487–1511, 2007.
- W. Jiang and M. Tanner. Gibbs posterior for model selection in high dimensional classification and data mining. *The Annals of Statistics*, 36:2207–2231, 2008.
- W. Jiang and B. Turnbull. The indirect method: Inference based on intermediate statistics A synthesis and examples. *Statistical Science*, 19:239–263, 2004.
- V. E. Johnson and D. Rossell. On the use of non-local prior prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society, Series B*, 72:143–170, 2010.
- V. E. Johnson and D. Rossell. Bayesian model selection in high dimensional settings. *Journal of the American Statistical Association*, 107:649–660, 2012.
- K. Kato. Quasi-bayesian analysis of nonparametric instrumental variables models. *The Annals of Statistics*, 41:2359–2390, 2013.
- J. Y. Kim. Limited information likelihood and Bayesian analysis. *Journal of Econometrics*, 107:175–193, 2002.
- Y. Kitamura and T. Otsu. Bayesian analysis of moment condition models using nonparametric priors. Mimeo. 2011.
- Y. Kitamura and M. Stutzer. An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65:861–874, 1997.

- A. Kottas and A. E. Gelfand. Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, 96:1458–1468, 2001.
- G. Kundhi and P. Rilstone. Edgeworth expansions for GEL estimators. *Journal of Multivariate Analysis*, 106:118–146, 2012.
- G. Kundhi and R. Rilstone. Edgeworth and saddlepoint expansions for nonlinear estimators. *Econometric Theory*, 29:1057–1078, 2013.
- T. Lancaster and S. J. Jun. Bayesian quantile regression methods. *Journal of Applied Econometrics*, 25:287–307, 2010.
- N. A. Lazar. Bayesian empirical likelihood. *Biometrika*, 90:319–326, 2003.
- C. Leng and C. Y. Tang. Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika*, 99:703–716, 2012.
- G. Li, H. Peng, and L. Zhu. Nonconcave penalized  $M$ -estimation with a diverging number of parameters. *Statistica Sinica*, 21:391–419, 2011.
- G. Li, H. Lian, S. Feng, and L. Zhu. Automatic variable selection for longitudinal generalized linear models. *Computational Statistics and Data Analysis*, 61:174–186, 2013.
- Q. Li, R. Xi, and N. Lin. Bayesian regularized quantile regression. *Bayesian Analysis*, 5:533–556, 2010.
- F. Liang, R. Paulo, G. Molina, M. Clyde, and J. O. Berger. Mixture of  $g$ -priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423, 2008.
- F. Liang, Q. Song, and K. Yu. Bayesian subset modeling for high-dimensional generalized linear models. *Journal of the American Statistical Association*, 108:589–606, 2013.
- K-Y Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.
- Y. Liao and W. Jiang. Bayesian analysis in moment inequality models. *The Annals of Statistics*, 38:275–316, 2010.
- Y. Liao and W. Jiang. Posterior consistency of nonparametric conditional moment restricted models. *The Annals of Statistics*, 39:3003–3031, 2011.

- J. Marin, N. S. Pillai, C. P. Robert, and J. Rousseau. Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society, Series B*, To appear, 2013.
- E. Moreno, F. J. Girón, and G. Casella. Consistency of objective Bayes factors as the model dimension grows. *The Annals of Statistics*, 38:1937–1952, 2010.
- W. K. Newey. Efficient semiparametric estimation via moment restrictions. *Econometrica*, 72:1877–1897, 2004.
- W. K. Newey and R. J. Smith. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72:219–255, 2004.
- A. B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249, 1988.
- S. Portnoy. Asymptotic behavior of  $M$ -estimators of  $p$  regression parameters when  $p^2/n$  is large. I. Consistency. *The Annals of Statistics*, 12:1298–1309, 1984.
- S. Portnoy. Asymptotic behavior of  $M$ -estimators of  $p$  regression parameters when  $p^2/n$  is large. II. Normal approximation. *The Annals of Statistics*, 13:1403–1417, 1985.
- J. Qin and J. Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22:300–325, 1994.
- A. Qu, B. G. Lindsay, and B. Li. Improving generalized estimating equations using quadratic inference functions. *Biometrika*, 87:823–836, 2000.
- P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39:731–771, 2011.
- S. M. Schennach. Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92:31–46, 2005.
- S. M. Schennach. Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, 35:634–672, 2007.
- J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38:2587–2619, 2010.
- M. Smith and R. Kohn. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75:317–343, 1996.



- M. E. Stokes, C. S. Davis, and G. G. Koch. *Categorical Data Analysis using the SAS System*. SAS Institute Inc., Cary, NC, 2000.
- V. W. van der Vaart and J. A. Wellner. *Weak converge and empirical processes: with applications to statistics*. Springer, New York, 1996.
- L. Wang. GEE analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics*, 39:389–417, 2011.
- L. Wang and A. Qu. Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *Journal of the Royal Statistical Society, Series B*, 71:177–190, 2009.
- L. Wang, J. Zhou, and A. Qu. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics*, 68:353–360, 2012.
- S. Wang, L. Qian, and R. J. Carroll. Generalized empirical likelihood methods for analyzing longitudinal data. *Biometrika*, 97:79–93, 2010.
- Y. Yang and X. He. Bayesian empirical likelihood for quantile regression. *The Annals of Statistics*, 40:1102–1131, 2012.
- G. Yin. Bayesian generalized method of moments. *Bayesian Analysis*, 4:191–208, 2009.
- K. Yu and R. A. Moyeed. Bayesian quantile regression. *Statistics and Probability Letters*, 54:437–447, 2001.
- A. Yuan and B. Clarke. Asymptotic normality of the posterior given a statistic. *The Canadian Journal of Statistics*, 32:119–137, 2004.